Addressing non-adherence in cluster randomised trials using instrumental variable-based methods

Schadrac Christin Agbla

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of London

Department of Medical Statistics

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

FEBRUARY 2020

Funded by the UK Economic and Social Research Council (ESRC), Grant code: ES/J500021/1.

Declaration

I declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been acknowledged in the thesis.

Signature

Schadrac Christin Agbla

Date: February 2020

Abstract

Randomised trials are viewed as the gold standard for evaluating interventions. Depending on the intervention as well as other logistical factors, individuals or group of individuals may be randomised. The former is known as individual randomised controlled trials (RCTs) and the latter as cluster randomised trials (CRTs). CRTs offer advantages such as administrative convenience and reduction of contamination between trial groups but analysis is more complex than that for RCTs, because of the correlations between participants in the same cluster. When non-adherence to treatment occurs in the sense that some participants do not receive the randomly assigned treatment, confounding may exist as there may be common factors influencing treatment received and outcome. Consequently, the intention-to-treat approach, which compares outcomes between the groups as randomised, assesses the effect of being randomised to treatment rather than the causal treatment effect (effect of receiving the treatment).

Ad-hoc methods often used to attempt to estimate the causal effect of treatment received such as per-protocol (PP) and as-treated (AT) approaches are likely to provide biased estimates because the assumptions necessary for those approaches to be unbiased are in general implausible. There exists extensive literature on estimating causal treatment effects from RCTs with non-adherence, but not as much for CRTs. Instrumental variables (IV) methods have the advantage, over other causal methods, of accommodating settings where there are unmeasured confounders when making causal inference.

This thesis contributes to the literature on the estimation of causal treatment effects in CRTs where there is non-adherence to treatment and focuses on IV-based methods. I first ascertained the current practice of reporting and addressing non-adherence when causal treatment effects are of interest in CRTs via a systematic review of 123 CRT reports. Non-adherence was reported in about half of the CRTs,

of which a third were interested in the causal treatment effect. All of the reviewed CRTs that reported adherence-adjusted estimates performed either PP or AT, without discussing the plausibility of the very strong assumptions necessary for such analyses to result in unbiased causal treatment estimates. No study estimated the local average treatment effect (LATE), that is the average treatment effect on those that would comply with the random allocated treatment, or any other appropriate statistical methods for unbiased causal estimation.

In many clinical settings, the relevant causal question is whether treatment has an effect among those who are willing to take it, which would be quantified by the LATE. Hence the thesis focuses on this estimand, starting with an introduction and assessment of the performance of IV-based methods for estimating LATE at either cluster level (CL) or individual level (IL) through simulations under the required identification assumptions for LATE. I also perform sensitivity analyses for IL-LATE estimation and illustrate those methods using two real CRTs. The methods include two-stage least squares (TSLS) based on CL outcome summaries and the Wald estimator with the Schochet-Chiang standard error to estimate CL-LATE, and the Wald estimator, TSLS with robust cluster standard errors, TSLS with Moulton's standard errors and the Bayesian multilevel mixture modelling for estimating IL-LATE. I conduct extensive simulations and illustrate the methods using real CRTs data. I demonstrate that TSLS is attractive for the estimation of CL-LATE and IL-LATE but is inefficient. This inefficiency may be reduced through covariate adjustment. The Bayesian multilevel mixture modelling is also attractive due to its flexibility and performs well particularly when non-adherence is at the individual level and the intracluster correlation coefficient for outcome is large. Stata and R codes are provided to facilitate implementation by trial investigators. I conclude by making some recommendations about how to estimate CL-LATE and IL-LATE to improve the quality of analysis when estimating causal treatment effects in the presence of non-adherence in CRTs.

Acknowledgements

I am deeply grateful to my supervisor Dr Karla Diaz-Ordaz and associate supervisor Prof Bianca DeStavola for their commitment all these years. I was very fortunate to have them involved in my PhD and to receive their guidance and critical insights.

I would like to thank Dr Elizabeth Williamson aka Fizz for giving me access to the TXT4FLUJAB trial data used as a motivating example and also thank the UK Economic and Social Research Council (ESRC) via the Bloomsbury Doctoral Training Centre for providing the financial support. I cannot forget Jenny Fleming and Lauren Dalton, who have been extremely awesome in dealing with the administrative issues and have made my years at the school so pleasant to the point that I could carry on with my PhD forever. Unfortunately, I have to move to a next step in my academic life. Jenny and Lauren, thank you!

My sincere gratitude also is towards my family and friends at and outside LSHTM for their continuous support and encouragement. I prefer not to name them to avoid omitting any of them as the list is extremely long and may double the number of pages of this document. However, with no intention to offend the rest, I would like to thank my officemates Anower ("our daddy"), Kleio ("the crazy superwoman"), Simon ("the robot"), Tom ("Tom Cruise's body double") and Ollie ("my little geek brother") with whom I shared the same heat, cold and water drip noise throughout my PhD.

Finally, I thank God Almighty for taking care of me and strengthening me from the day I was born until today.

Contents

List of Tables

List of Figures

Abbreviations

adCL	Adjusted cluster-level
AT	As-treated
ATE	Average treatment effect
ATT	Average treatment effect on the treated
Bern	Bernoulli distribution
Bin	Binomial distribution
BMM	Bayesian multilevel mixture
CACE	Complier average causal effect
CL	Cluster level
Cov	Covariance
CRA	Complete records analysis
CRT	Cluster randomised trial
CS	Cluster size
DAG	Direct Acyclic Graph
EM	Expectation-maximization
$F-$statistic	Fisher statistic
GP	General practice
HW	Huber-White
HWR	Huber-White-Rogers
ICC	Intracluster correlation coefficient
IL	Individual level
ITT	Intention-to-treat
IV	Instrumental variable
LATE	Local average treatment effect
MCE	Monte Carlo Error
MCMC	Markov Chain Monte Carlo
MV	Minimum variance
OLS	Ordinary least squares
PO	Potential outcome

Poi	Poisson distribution
PP	Per-protocol
PS	Principal stratification
RCT	Randomised controlled trial
RSS	Residual sum of squares
SD	Standard deviation
SE	Standard error
SSDF	Small sample degrees of freedom
TSLS	Two-stage least squares
TSS	Total sum of squares
$t-$statistic	Student statistic
unCL	Unadjusted cluster-level
Var	Variance
WLS	Weighted least square
WTLS	Weighted two-stage least square

Dissemination

Published research papers

- Agbla, S.C. and DiazOrdaz, K., 2018. Reporting non-adherence in cluster randomised trials: A systematic review. Clinical Trials, 15(3), pp.294-304.
- Agbla, S.C., De Stavola, B. and DiazOrdaz, K., 2019. Estimating cluster-level local average treatment effects in cluster randomised trials with non-adherence. Statistical Methods in Medical Research, p.0962280219849613.

Oral presentations

- Causal treatment effects in cluster randomised trials: estimation by cluster-level instrumental variable methods. Presented at the Current development in cluster randomised trials and stepped wedge designs Conference in London, Queen Mary University. 30 November 2017.
- Addressing non-adherence in cluster randomised trials: estimation by cluster-level instrumental variable methods. Presented at the 44th Young Statisticians' Meeting in Oxford, Department of Statistics. 30-31 July 2018.
- Estimation of LATE in cluster randomised trials. Presented at the Symposium of the 50 years of the Masters in Medical Statistics, London School of Hygiene & Tropical Medicine. 12 April 2019.

Poster presentation

- Estimation of causal treatment effect at cluster level in cluster randomised trials with non-adherence. Presented at the Research Degrees Poster Day at the London School of Hygiene & Tropical Medicine. 15 March 2018.

Chapter 1.

Introduction

This chapter describes randomised trials in general and introduces cluster randomised trials. I clarify what is meant by non-adherence to treatment and highlight its implications when investigators are interested in causal treatment effects. I use the terms "treatment" and "intervention" interchangeably. Without formal notation, I provide a brief review of the causal estimands often used in the biostatistics literature. Finally, I introduce the trials motivating this work and delineate the scope of the thesis.

1.1 Rationale for randomised trials

Randomised trials are studies where experimental units are randomly assigned to either control treatments (referred to as control groups) or active treatments (referred to as active groups). Here, I focus on two-arm randomised controlled trials. The random allocation of experimental units to the control or active group, referred to as randomisation, aims to prevent possible biases such as confounding bias that may cause systematic differences between trial groups [1, 2]. Randomisation allows experimental units to have a known probability, equal or unequal, of being assigned to either treatment while the treatment assigned to each experimental unit cannot be predicted [2]. Thus, the treatment assignment mechanism is unconfounded, that is, there are no measured or unmeasured variables that may influence the treatment allocation and are independently associated with the outcomes of interest. The unconfoundedness of treatment assignment is a key feature of randomised trials as it allows us to attribute any observed average difference in the outcome variable across trial groups to the assignment of the active treatment.

Randomised trials are viewed as the gold standard for testing new interventions in many disciplines such as public health, epidemiology and social sciences. Through

randomisation, we aim to balance the distribution of measured and unmeasured variables across trial groups, making individuals belonging to the control and active treatment group exchangeable in the sense that they would experience the same outcome probability to those in the other group, if assigned to it.

1.2 Individual and cluster randomised trials

There exists two types of randomised controlled trials in terms of the nature of the experimental units involved, that is, individual randomised controlled trials (RCTs) and cluster randomised trials (CRTs). RCTs are trials where the units of randomisation are individuals such as patients and pupils. CRTs, however, are experiments where groups of individual units such as those belonging to the same hospital, school and community, or batches of rats are allocated to either the control or active treatment. CRTs have been increasingly used to assess complex public health, education or economic interventions that target groups of individuals. CRTs offer some practical advantages over RCTs such as administrative convenience, reduction of contamination between trial groups and improved adherence to treatment [3–5]. Despite its practical advantages, there are statistical complexities that may arise in CRTs because of the dependence of the data that has to be accounted for.

However, selection bias may potentially occur in some CRTs where individual units are recruited after the clusters have been randomly allocated to a treatment group or when consent is needed prior to treatment [6,7]. This selection bias can be attenuated by either identifying participants and obtaining their consent prior to clusters' randomisation or allowing a third-party to blindly identify and select participants into randomised clusters [6].

1.2.1 Intracluster correlation coefficient

The intracluster correlation coefficient (ICC) is a measure of how much of the total variance is accounted for by the between-cluster variation. The ICC represents the degree of similarity across individuals within the same cluster than those from other clusters [8]. The mathematical expression of the ICC denoted by ρ is

$$\rho = \sigma_{\text{B}}^2/\sigma^2 = \sigma_{\text{B}}^2/(\sigma_{\text{B}}^2 + \sigma_{\text{W}}^2), \tag{1.1}$$

where σ_B^2 is the between-cluster variance, σ_W^2 the within-cluster variance and σ^2 the total variance.

1.2.2 Levels of analysis

Depending on the aims of the trial, data from CRTs can be analysed at the cluster or individual level. For example, cluster-level analyses may be suitable for CRTs where investigators are interested in assessing the effect, of general practices (GPs) sending reminder text messages to their patients, on the proportion of vaccinated patients rather than the effect on the probability of a patient getting vaccinated. The latter requires the analysis to be done at the individual level.

The analysis at the cluster level consists of first computing cluster-level summary statistics of the outcome variable (for instance, means for continuous variables and proportions for binary variables) and then performing the planned statistical comparisons on the cluster-level summaries such as $t-$test and linear regression [8]. Allowance for differential cluster sizes or varying within-cluster heterogeneity can be made by weighing the contributions of the various clusters to the group comparisons. For instance, this can be achieved with weighted regression. In the cluster-level summaries analysis, the number of units contributing to the analysis is equal to the number of clusters.

On the other hand, the analysis at the individual level uses data as collected (*i.e.* preserving its hierarchical structure). Therefore, the within-cluster correlations need to be accounted for to ensure valid inference. This can be achieved through mixed effects regression (linear, Poisson or logistic for instance, depending on the type of outcome variable) or generalised estimating equations (GEE) [8].

The level of analysis is determined by the unit of inference [9] and should be clarified in the protocol at the trial design stage [10]. Analysis at the cluster level may be preferable when there is small number of clusters. In such settings, inference at the individual level may not be as reliable as the between-cluster variance may be poorly estimated [8].

1.3 Non-adherence in randomised trials

By non-adherence to treatment (also referred to as non-compliance), I mean deviations from the assigned treatment such that some participants randomised to the control group receive the active treatment and/or some of those allocated to the active treatment receive the control treatment. Note that loss of follow up is viewed as a missing data problem rather than as non-adherence to treatment. I use "non-adherence to treatment" or simply "non-adherence" or "non-compliance", interchangeably.

Non-adherence often occurs in both RCTs and CRTs. A systematic review of RCTs published in 2008 in BMJ, New England Journal of Medicine, the Journal of the American Medical Association and The Lancet found that 98 out of 100 RCTs were reportedly affected by non-adherence [11]. Regarding CRTs, non-adherence was reported in about 24% and 13% out of 152 published CRTs (searched from the Cochrane Controlled Trials Register) and 47 unpublished CRTs (searched fron relevant conference proceedings and the UK National Research Register) between 1997-2000, respectively [12].

Interventions in RCTs are administered directly to individuals and therefore, non-adherence is related only to the individuals. In CRTs, however, because of the hierarchical nature of the design, interventions can be implemented either at the cluster level, or individual level, or even at both cluster and individual levels [13]. Therefore, non-adherence may occur at least at one of cluster and individual level, depending on the nature of the intervention. Non-adherence is considered to be at the cluster level if the treatment received was different from that assigned for all the participants within clusters, and to be at the individual level if the treatment received differed from the allocated treatment on an individual basis within the same cluster. Schochet et al. [13] introduced settings where non-adherence occurs at both cluster and individual levels and interact, adding complexity to the analysis. I only consider setting where non-adherence does not interact at both levels.

When the control group is not allowed to have access to the active treatment, we

have "one-sided non-adherence", whereas there is "two-sided non-adherence" when both control and active groups are subject to non-adherence.

The recommended approach for analysing randomised trial data is the so-called intention-to-treat (ITT) analysis [14], that is, the comparison of experimental units' outcomes between trial groups, regardless the treatment actually received [15, 16]. However, non-adherence in randomised trials has some implications regarding the interpretation of the ITT results. This is covered in the next section.

1.4 Brief review of causal estimands

This section presents a brief review of causal estimands (quantities of interest in a population that we want to estimate using a sample from that population) and some approaches often used in the causal inference literature to measure the effect of a treatment in the presence of non-adherence.

1.4.1 Intention-to-treat

Two general causal questions can be asked in randomised trials when there is non-adherence. The first and most common question addressed is: "what is the causal effect on the outcome of offering this treatment versus an alternative one?". The second is "what is the causal effect on the outcome of actually receiving versus not-receiving the treatment?". The first question is answered via an ITT analysis whereas addressing the second one generally requires a more elaborate approach. In randomised trials where all experimental units receive their allocated treatment, the two causal questions are equivalent and the ITT approach is appropriate.

In settings where non-adherence occurs without using treatments outside those in the trial (rescue medication for example), the ITT effect is closer to the null than the causal effect [17]. Thus, when considering negative outcomes such as side effects, adverse events or mortality, ITT effect may make a treatment appear less harmful [17,18]. When there is non-adherence, the ITT effect assesses the benefit of offering the treatment compared to an alternative treatment (so-called "effectiveness") instead of the benefit of actually receiving the offered treatment or intervention (also known as "efficacy").

1.4.2 Per-protocol and as-treated approaches

There are two popular approaches attempting to estimate the effect of actually receiving the treatment in the presence of non-adherence: the per-protocol (PP) and as-treated (AT) analyses [11]. PP seeks to estimate the causal treatment effect by restricting the analysis only to experimental units who received their assigned treatment. If there exist pre-treatment factors that influence both the treatment received and the outcome, the subgroup of experimental units who received their assigned treatment is not guaranteed to have similar baseline characteristics in the control and active groups. Thus, PP may be subject to selection bias and also leads to a reduction of statistical power because of the exclusion of those who did not receive the allocated treatment [18, 19].

The AT analysis is done by comparing experimental units according to the treatment received regardless of their treatment assignment [20]. The AT approach also suffers from potential bias because the random allocation of treatment is substituted with the actual treatment received which may depend on some factors that also affect the outcome in those who do not comply. The key feature of random assignment which is to guard against confounding and ensure groups' comparability may be altered, making the interpretation of the results difficult [21].

PP and AT analyses lead to valid causal estimates of treatment effect if the subgroups being compared are exchangeable or conditionally (*i.e.* all the variables associated with the treatment received and the outcome are adjusted for) exchangeable [17]. It is very unlikely that the exchangeability assumption is met in practice. Treatment received after randomisation may likely not be random but may be influenced by factors affecting both treatment received and outcome. These confounding factors may also not be trivial to measure for a subsequent adjustment. However, some design such as double-blinding may make the unbiasedness of PP and AT estimates of causal treatment effect more plausible [18].

1.4.3 Formal causal estimands

It is helpful to introduce some formality in order to deal with the potential bias affecting PP and AT analyses. Statistical methods relying on more realistic assumptions than that of exchangeability or exchangeability conditional on observed pre-treatment variables have been developed and target estimands such as the average causal or treatment effect (ATE) and the average treatment effect among treated (ATT) have been proposed [17, 22, 23].

Another estimand is the local average treatment effect (LATE), that is the average treatment effect among the subgroup of participants that would receive the random treatment they would have been assigned to. LATE is sometimes referred to as complier average causal effect (CACE) in a setting of binary treatment assignment and binary treatment received [24].

These estimands are formally defined in terms of potential outcomes [25, 26] which are presented in chapters 4 and 6. However, their identification (ability to estimate those estimands from the data) like any causal estimands involves assumptions that are often untestable [26]. The choice of causal estimands is partly driven by the assumptions that one is willing to make and partly driven by the scientific question and the data available.

The random treatment assignment can be used as an instrumental variable (IV), that is a variable fulfilling the following criteria: – (i) independent of measured and unmeasured confounders of the treatment received-outcome relationship, – (ii) associated with treatment received, and – (iii) does not directly affect outcome, except through the treatment received (known as *exclusion-restriction* (ER)). Let A be any measured pre-randomisation covariate, L any post-randomisation covariate, U any unmeasured covariate, Z the random treatment assignment, D the treatment received and Y the outcome variable of interest. The relationship between A, L, U, Z, D and Y, where Z is an IV, can be represented as shown in Figure 1.1.

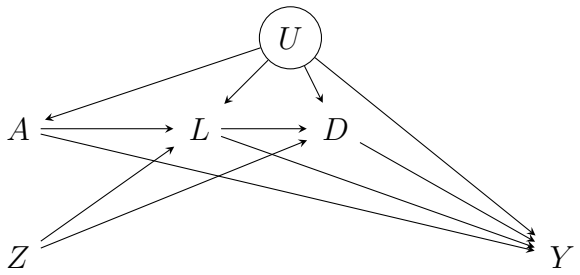


Figure 1.1: Diagram summarising the relationship between random treatment assignment (Z), treatment received (D), measured covariates (A and L), unmeasured covariate (U) and outcome (Y)

The bounds (minimum and maximum values) of LATE are identifiable when an IV is available. The LATE is point-identified with the additional assumption that the Z–D relationship is monotonic *i.e.* all experimental units if assigned to the active group would all either be encouraged or all discouraged to receive the active treatment (known as *monotonicity*) [24, 27, 28]. An IV analysis accounts for measured and unmeasured confounding, without the need to adjust for extra confounders [23].

In many clinical settings, the relevant causal question is whether treatment has an effect among those who are willing to take it, which would be quantified by the LATE. Moreover, the required identification assumptions for LATE are easily met in randomised trials. Hence the causal estimand of interest throughout this thesis is the LATE. A summary of the assumptions for the identification LATE is shown in Box 1.4.3. Chapters 4 and 6 formally present in detail the identification assumptions for LATE.

Box 1.4.3: Identification assumptions for LATE
A key idea is that of potential outcomes, *i.e.* the outcome that would have been observed had the randomised allocation been different. Estimands are then defined in terms of these potential outcomes, with their definitions varying according to the populations to which they refer. Likewise, the potential treatment received is the treatment that individuals/clusters would have received had their randomised allocation been different. Assuming all-or-nothing adherence *i.e.* setting where adherence is defined as a binary variable [29], the most common assumptions are:

i. **Stable Unit Treatment Value Assumption** (SUTVA): the potential outcomes of the i-th individual are unrelated to the treatment status of all other individuals

(known as *no interference*). In addition, *consistency* is assumed *i.e.* for those who actually received treatment level z, the observed outcome is the potential outcome corresponding to that level of treatment. In CRTs, SUTVA is unlikely to hold. Instead, we may assume that *no interference* holds at the cluster level, *i.e.* the potential outcome of an individual is unrelated to the treatment status of individuals in different clusters, but may depend on those within the same cluster [13, 30].

ii. **Ignorability of the treatment assignment**: Randomised allocation is independent of unmeasured confounders (conditional on measured covariates) and the potential outcomes.

iii. **Instrument relevance**: The random allocation predicts treatment received.

iv. **Exclusion restriction** (ER): The random allocation cannot affect the outcomes directly.

v. **Monotonicity**: There are no *defiers*, *i.e.* individuals who receive treatment only if they are not randomised to it. Generalisations of these assumptions to CRTs settings as in Schochet and Chiang [13] have been presented in Chapter 4.

1.4.4 Methods for estimating LATE

Methods such as the Wald estimator (ratio of the effect of the IV on the outcome to the effect of the IV on the treatment received) [13,31], the two-stage least square (TSLS) estimation [28] and mixture modelling [32,33] can be used to estimate LATE, assuming that the IV assumptions and monotonicity hold in addition to the technical assumptions of SUTVA (*consistency* and *no interference*). Multilevel mixture modelling can be implemented within both a frequentist and Bayesian framework. These methods are detailed in Chapters 4, 6 and 9.

The use of mixture modelling involves principal stratification [34] which I briefly present in Box 1.4.4 below.

Box 1.4.4: Principal stratification

1. **Instrumental variables** (IV): Under assumptions (i)-(v), Angrist et al. [24] showed that LATE is the ITT effect among those that would receive the active treatment if they are assigned to it and can be estimated using IV-based method. LATE is usually estimated using TSLS but the Wald estimator can also be used

[13, 31]. To account for the clustering, it has been suggested for instance to use TSLS estimation using Huber-White variance estimator [35].

2. **Principal stratification** [34]: Under assumptions (i)-(v), each individual may be grouped into a compliance (or adherence) *principal stratum*, which is a latent class, and can be thought of as a baseline covariate defined based on two potential outcomes for treatment received.
   (a) Never-takers receive no active treatment, regardless of their randomised treatment;
   (b) Compliers receive the active treatment only if they are randomised to it;
   (c) Always-takers receive the active treatment, regardless of their randomised treatment.

The principal stratification is based on a mixture of distributions across adherence classes, hence the use of mixture modelling to find out the adherence classes from the observed data and then estimate LATE. Extensions to CRTs are possible, by using multilevel mixture models, in either a Bayesian [34] or likelihood approaches [30].

1.5 Illustrative examples

I now briefly introduce the TXT4FLUJAB and the OPERA CRTs that I use to illustrate some of the analysis approaches. Those examples are complementary to a simulation study carried out in this thesis. The simulations do not particularly follow the non-adherence patterns nor the CRT sizes (number of clusters and individual units within clusters) in either illustrative examples.

1.5.1 The TXT4FLUJAB trial

The TXT4FLUJAB trial was a non-blinded and two-sided non-adherence CRT implemented in the United Kingdom, aiming at estimating the effect of text messages reminding patients at risk of chronic conditions to get the influenza vaccination during the 2013 influenza season [36]. General practices (GPs) were stratified by the type of software used for text messaging. There were three strata, each of which characterized by either of the following messaging software: – (i) "CPRD", – (ii) "ResearchOne and SystmOne", and – (iii) "EMIS", "Vision", or "Immform". GPs were randomised to either standard care (control group, 79 GPs and 51 136 patients) or a text messaging campaign (active group, 77 GPs and 51 121 patients).

GPs represent the clusters and patients were the individual units.

Adherence was binary and measured at the patient level. Patients whose GPs were allocated to the active group are said to adhere if they received the reminder text messages; whereas, those whose GPs were allocated to the control group are said to adhere if they did not receive any reminder text message.

GPs were the unit of analysis and the outcome of interest was the proportions of influenza vaccine uptake at the GPs level. The trial investigators were interested in the following causal questions: – (i) a question motivating the ITT analysis: "what is the causal effect on influenza vaccine uptake at the GPs level, of sending reminder text messages to patients at risk of chronic conditions to get the influenza vaccination compared to not sending any reminder text message?", and – (ii) a question motivating the LATE estimation: "what is the causal effect on influenza vaccine uptake at the GPs level, of patients at risk of chronic conditions actually receiving reminder text messages to get the influenza vaccination compared to not receiving any reminder text message?". These causal treatment effects at the GPs level were expressed as a mean risk percentage difference.

1.5.2 The OPERA trial

The OPERA trial was a non-blinded CRT conducted in Warwick and London between 2008-2010. It was carried out in 78 care-homes (35 allocated to the active group and 43 to the control group). The intervention was a complex programme which involved training on depression awareness for care home staff, 45 minutes physiotherapist-led group exercise sessions for residents (delivered twice a week) and a whole home component designed to motivate residents to do more daily physical activity. Staff of care-homes in the control group only received the training on depression awareness [37]. In total, 900 residents were enrolled (498 in the control group and 402 in the active group).

No assessment of causal treatment effects was planned by the investigators and therefore, there was no definition of adherence. However, for illustration purposes, we used a working definition of adherence as follows: residents in the active group

are considered to have adhered if they attended at least one group exercise session, whereas those in the control group adhered if they did not attend any group exercise sessions. On that basis, adherence was 100% in the control group, whereas 89% of residents in the active group adhered. Hence, the OPERA trial had one-sided non-adherence.

The outcome of interest is a secondary outcome defined by the trial protocol, that is, the Short Physical Performance Battery (SPPB) at 12 months since trial initiation [37]. SPPB is a score varying from 0 (worst performance) to 12 (best performance) and measures functional mobility. SPPB consists of three components of physical function: walking speed, standing balance and sit-to-stand performance [38]. SPPB was analysed as a continuous variable by the trial investigators [37].

The trial investigators were only interested in the ITT causal question that is: "what is the causal effect on SPPB at 12 months at the resident level, of offering the intervention as opposed to standard care?". The ITT analysis showed weak evidence of ITT effect [37]. I further investigate whether there may be a causal effect of the intervention and formulate the following question motivating the LATE estimation presented in chapter 8, that is "what is the causal effect at the cluster and individual level of receiving the intervention compared to standard care, on SPPB at 12 months?".

1.6 Thesis scope

Estimation of causal treatment effects in RCTs where there is non-adherence has been addressed extensively. However, for CRTs, extensions to accommodate clustering and non-adherence are available but the literature and applications are limited. For instance, methods listed in section 1.4.4 (Wald estimator, TSLS and Bayesian mixture modelling) are potential approaches for estimating LATE in CRTs where there is non-adherence but their performance has not been explored. This thesis aims to fill this gap by investigating the finite-sample performance of those methods where the analysis is at either the cluster or the individual level, allowing for covariate-adjustment and missing data and different approaches for estimating confi-

dence intervals, and thus contributing to improving good statistical analysis practice when estimating the causal treatment effects in CRTs where there is non-adherence. In addition, I extend existing sensitivity analysis strategies to CRT settings and illustrate their use in practical applications (see Chapter 9).

The focus is on IV-based methods and their applications because of reasons provided in section 1.4. The scientific question in clinical investigation is often one involving compliers. Thus, the estimand of interest is the LATE. The objectives of this research are as follows.

1. **To provide a description of the current practice of reporting and addressing non-adherence in CRTs:** The review by Eldridge *et al.* [12] gave an insight about the presence of non-adherence in CRTs and pointed out the lack of adequate statistical methods to account for clustering. However, this review included CRTs reports from 1997 to 2000. I investigate the current practice on the reporting of non-adherence and how non-adherence is addressed when the causal treatment effect is of interest. This is necessary to understand the extent of the problem and identify the gaps needed to be filled. I therefore conducted a systematic review reported in chapter 2.

2. **To assess the performance of different methods estimating LATE at the cluster level:** From the systematic review in chapter 2 emerged a need for appropriate statistical methods for estimating causal treatment effects in CRTs where there is non-adherence at the cluster and individual level. Therefore, I dedicate chapter 4 to examining the performance of TSLS estimation on cluster-level summaries to estimate cluster-level LATE through simulations in the presence of cluster-level or individual-level non-adherence. Application of TSLS and Wald estimations on the TXT4FLUJAB and the OPERA trials data are described in chapters 8.

3. **To assess the performance of different methods estimating LATE at the individual level:** This complements the previous objective in addressing the need for appropriate estimation methods for individual-level effects, in the

presence of cluster-level or individual-level non-adherence. I compare the results obtained from applying IV-based methods for estimating the individual-level LATE, using data from the TXT4FLUJAB and the OPERA trials in chapters 8. Sensitivity analyses using the OPERA trial are also conducted in chapter 9, to assess the robustness of individual-level LATE estimates when relaxing the ER assumption.

1.7 Thesis structure

The rest of the thesis is organised as follows. Chapter 2 presents a systematic review conducted to ascertain the extent of non-adherence in CRTs and the current practice in terms of addressing non-adherence when estimating causal treatment effects is of interest.

Chapter 3 introduces cluster-level summary approaches in CRTs. Chapter 4 focuses on cluster-level summary-based analyses, lays the assumptions required for the identification and estimation of LATE at the cluster level, and presents alternative statistical approaches to estimate cluster-level LATE. Chapter 5 presents a simulation study assessing the performance of estimation methods of cluster-level LATE presented in Chapter 4 and lessons about when and how these statistical approaches should be used, assuming all required assumptions for LATE identification and estimation hold.

Chapter 6 addresses the estimation of LATE at the individual level, introduces the required assumptions and a range of statistical methods that can be used at that end. Chapter 7 assesses, via simulation, the performance of individual-level LATE estimation.

Chapter 8 illustrates methods introduced in chapters 4 and 6 using the TXT4FLUJAB and the OPERA trials. Chapter 9 investigates sensitivity analyses, through applications using the OPERA trial only, the robustness of LATE estimates at the cluster and individual levels, when there is a departure from the ER assumption.

Finally, Chapter 10 summarises the findings, discusses the strengths and limitations

of IV-based methods covered in this thesis and proposes some recommendations for estimating LATE in CRTs, whether at the cluster level or at the individual level. This chapter also presents avenues of future work.

Chapter 2.

Systematic review on the reporting and addressing of non-adherence in CRTs

2.1 Introduction

Consolidated Standards of Reporting Trials (CONSORT) guidelines for RCTs and CRTs have been proposed to improve the reporting of trials [14, 39]. Investigators are encouraged to follow those guidelines when conducting and reporting trials. However, despite the CONSORT check-list [14, 39] requesting explicitly to report numbers assigned, on treatment and analysed, previous systematic reviews found that adherence to treatment is often under-reported, and when reported, sufficient detail on how adherence was defined is often not included [5, 11]. CRTs are more complex to run, analyse and report than RCTs, and appropriate statistical methods accounting for the clustering in the data need to be used [3, 40]. This applies also to methods used to estimate causal treatment effects. Systematic review studies have been conducted to investigate the reporting and analysis practices in addressing non-adherence in RCTs [11, 41] but no similar study on our knowledge had been done for CRTs.

The current chapter establishes the prevalence of non-adherence and describes the methods used to obtain causal treatment effects. For this, I perform a secondary analysis of data originally extracted for a systematic review investigating the reporting and adjustment of missing data in CRTs [42]. I also propose guidelines for reporting adherence and conducting causal analysis of CRTs. The chapter is organised as follows. Section 2.2 describes the methods including the search, piloting, data extraction and analysis. Section 2.3 presents the results. Section 2.4 provides a summary of the findings and section 2.5 contrasts the results with previous studies.

This systematic review is published in Clinical Trials and a copy can be found in appendix A.3.

2.2 Methods

This includes the search strategy, inclusion criteria, piloting and validation, the data extraction and analysis of the data generated by the systematic review.

2.2.1 Search strategy and inclusion criteria

DiazOrdaz et al. [42] previously identified 526 CRT reports using a published electronic search strategy shown in Box 2.2.1, of which 188 were excluded because those reports did not meet the inclusion criteria based on the titles and abstracts. From the remaining 338 reports, they randomly excluded approximately a fifth (62 reports) and carried out the assessment of the full text on the remaining 276 reports. The random exclusion of a fifth of the reports eligible for full-text assessment was to reduce the workload without undermining the validity of the systematic review process. Out of the 276 reports assessed, 132 met the inclusion criteria for analysis.

The present review uses those 132 CRTs previously identified. The review protocol except the electronic search strategy is reported in Appendix A.1. Reports were eligible for inclusion if they were full reports of cluster randomised controlled trials, published in English in 2011. They were excluded if they were quasi-experimental, self-identified as pilot, feasibility, or preliminary studies; only reported cost-effectiveness or where no data at the individual level were collected. We also excluded crossover trials, where deviations from randomised treatment may include failure to follow the randomised sequence of treatments, and studies reporting only sub-samples of previously published CRTs data.

While the systematic review includes CRT reports published in 2011, it is important to note that the pre-print version of the updated CONSORT statement for CRTs [39] was available a year earlier and there were no new guidelines concerning the reporting and handling of non-adherence as compared to the previous version of the CONSORT statement [14]. Having said this, it is arguable that although

methods for estimating causal treatment effects in both RCTs and CRTs have been published prior to 2011 [24, 25, 27, 30, 43, 44], these methods were not well known or accepted in the clinical trials community. It is plausible that my findings no longer reflect current practices, as the recent focus on causal inference methodology within clinical trials, especially the addendum to the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) E9 guidelines [45], published in 2017, highlighting the importance of LATE estimands, may have had a positive impact in the current reporting and analysis practices.

Box 2.2.1: Electronic search strategy used by DiazOrdaz et al. [42]

#1 (randomized trial) AND (clinical trial) [All Fields]
#2 ((cluster randomization) OR (cluster randomisation) OR (cluster) OR (clustered) OR (clustering) OR (clusters) OR (group-randomized) OR (group-randomised) OR (randomisation unit) OR (randomization unit)) [All Fields]
#3 animal [All Fields]
4#1 AND #2
5#4 NOT #3
6 Feasibility [All Fields]
7 Pilot [All Fields]
8 Protocol [All Fields]
9 Review [All Fields]
#10 : #6 OR #7 OR #8 OR #9
#11 : #5 NOT #10
#12 : year [2011] (for scoping)
#11 : AND #12
(searched performed in PubMed on 18 June 2012)
("2011"[Date - Publication] : "2011"[Date - Publication])

2.2.2 Piloting and validation

Two researchers independently piloted a data extraction form using five randomly selected reports. This helped to identify extra relevant information to extract and to improve the study protocol. After updating the study protocol and the data extraction form, a random sample of fifteen reports was used for validation of the extraction procedures. In case of discrepancy, a final decision was made by consensus

and the appropriate information was recorded in the data extraction form. Once the team was satisfied with the extraction procedure, I performed the data extraction in the whole sample. When there was doubt or ambiguity, this was reviewed by the second extractor and a consensus was reached.

2.2.3 Data extraction

Data were extracted on one primary outcome per report, defined as that specified by the authors or, if not specified, the outcome used in sample size calculations. If no primary outcome was specified and no sample size calculation was reported, the first outcome presented in the abstract or manuscript was considered as primary. Information was collected on the type of cluster, the type of primary outcome (binary, continuous, categorical), whether a harm outcome was investigated [46], and the type of intervention given in the control arm (placebo, standard care or active). Information on the level of adherence (cluster-level or individual-level) was also recorded. Non-adherence was considered to be at the cluster level if the treatment received was different from that assigned for all the participants within clusters, and it was considered to be at the individual level if the treatment received differed from the allocated treatment on an individual basis within the same cluster.

Additionally, data on total number of clusters and individuals randomised and analysed were extracted as well as numbers of clusters and individuals receiving the allocated treatment. We defined treatment non-adherence as discrepancy between the allocated course of treatment and the actual treatment received [11]. Descriptions of treatment adherence, including intra-cluster correlation coefficient for treatment adherence [30], were also recorded, when reported. I also recorded information on adherence-adjusted analyses and whether clustering was accounted for.

I adapted the definitions by Dodd et al. [11] and extracted data about the duration of the intervention. A "one-off" intervention is defined as that which is received at a single time point, e.g. a surgery. A "short-term" intervention is defined as an intervention implemented at different time points over a short period; for example, five training sessions on the importance of breastfeeding over one week. Any other

recurrent intervention over an extended period of time was classified as a "long-term" intervention.

2.2.4 Analysis

Simple analyses were performed to describe the frequency of adherence-reporting and the reported methods used to adjust for non-adherence. I provide the median (and the first and third quartiles) of the number of clusters and individuals that are randomised, the number of clusters and individuals on treatment and the number of clusters and individuals included in the analysis.

For the percentage of non-adherence, I used the author-reported non-adherence when this was reported numerically. If not, I calculated the percentage of non-adherence for each study, from the data provided in the manuscript (the ratio between "off allocated treatment" participants to the total number randomised).

2.3 Results

After excluding 7 reports that used only sub-samples of CRTs data and 2 crossover trials, our final sample included 123 CRTs. See the Flow Chart, Figure 2.1. During the validation phase, the two extractors had an initial agreement of 93%, ultimately achieving consensus by discussion.

2.3.1 Trial characteristics

Trial characteristics are shown in Table 2.1. Interventions were mainly concerned with changing healthcare practices (63 trials, 51%), educational practices (27 trials, 22%) or lifestyle (25 trials, 20%). In most trials, standard care was used as the control intervention (96 trials, 76%). The primary outcome was either continuous (65 trials, 53%) or binary (57 trials, 46%), with one exception (multi-category). Adverse events were investigated in 12 trials (10%).

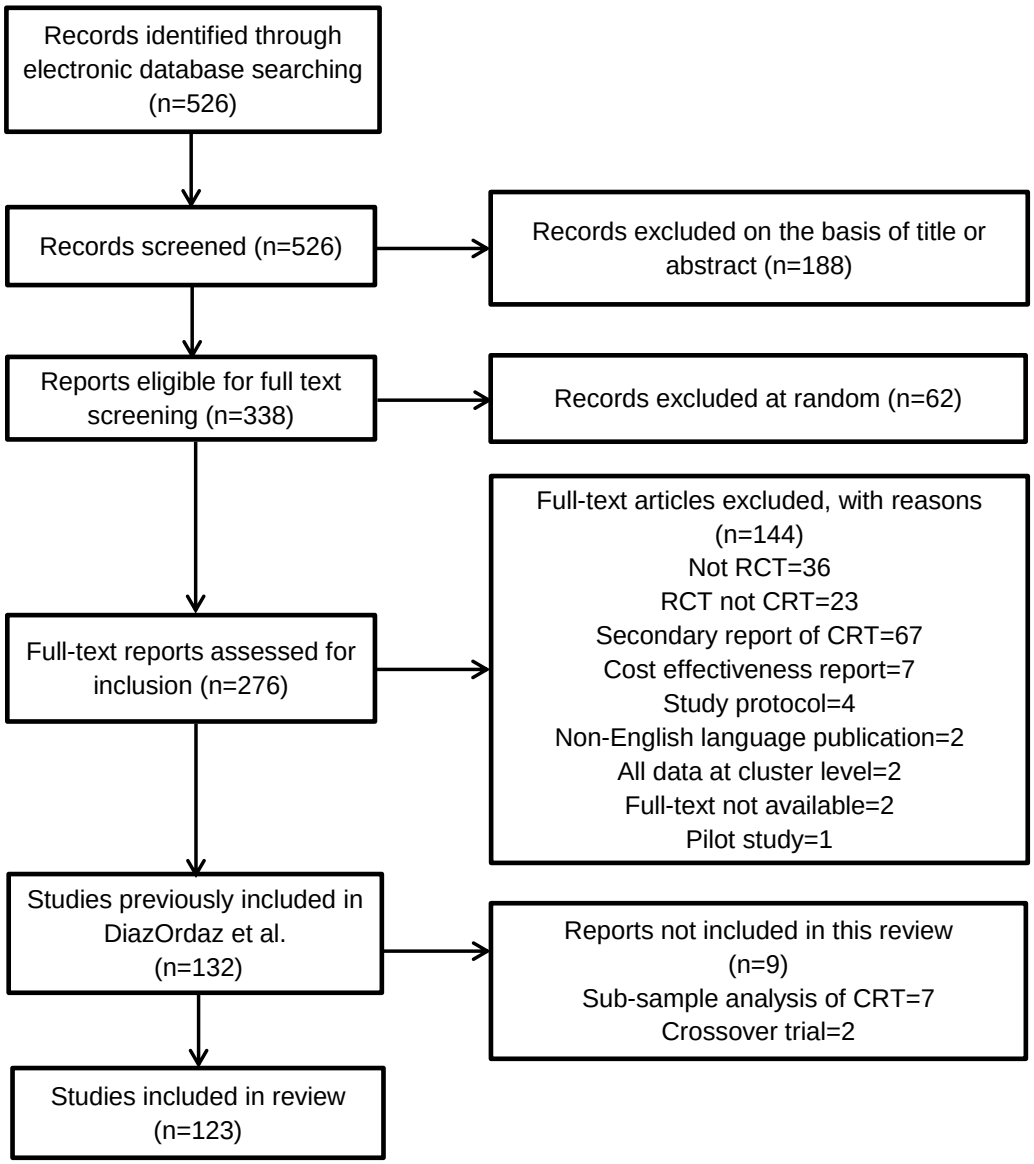


Figure 2.1: Flow diagram of the identification process for the sample of 123 CRTs included in this review

The intervention was implemented exclusively at the cluster level in 65 trials (53%) and at the individual level in 58 trials (47%). Long-term interventions were the most common (83 trials, 68%), followed by short-term interventions (35 trials, 28%). The majority of the studies were two-arm trials (106 trials, 86%). The median (1st–3rd quartiles) number of clusters randomised in each trial arm was 12 (7–24) and the number of clusters per trial arm ranged from 2 to 199. The number of individuals per cluster had a median (1st–3rd quartiles) of 27 (10–65) with a range of 2 to 14350.

Table 2.1: Characteristics of the CRTs included in this review

Characteristics	**Included trials (123 CRTs)**
Type of intervention, n (%)	
Healthcare practice	63 (51.2)
Lifestyle changes	25 (20.3)
Educational	27 (22.0)
New drug	5 (4.1)
Vaccination/screening	3 (2.4)
Type of control intervention, n (%)	
Standard practice	94 (76.4)
Active control	27 (22.0)
Placebo	2 (1.6)
Primary outcome, n (%)	
Continuous	65 (52.8)
Binary	57 (46.4)
Categorical	1 (0.8)
Investigation of adverse events, n (%)	12 (9.8)
Number of trial arms, n (%)	
2	106 (86.2)
3-4	17 (13.8)
Level of intervention, n (%)	
Cluster level	65 (52.8)
Individual level	58 (47.2)
Unit of analysis, n (%)	
Clusters	6 (4.9)
Individuals	117 (95.1)
Length of intervention, n (%)	
One-off	5 (4.1)
Short term	35 (28.4)
Long term	83 (67.5)
Clusters randomised per arm, Median (1st-3rd quartiles)	12 (7-24)
Range	2-199
Cluster size, Median (1st-3rd quartiles)[a]	27 (10-65)
Range[a]	2-14350
Primary analysis population, n (%)	
Intention-to-treat	119 (96.8)
Per protocol/as treated	4 (3.2)

[a] Based on the average number of individuals per cluster reported in each trials.

ITT analysis was done as primary analysis in 119 trials (97%), with the remaining 4 trials (3.2%) using PP or AT analysis. Only 6 trials (5%) used cluster-level analysis (primary outcome defined at the cluster level) while the remaining 117 trials use individual-level analysis. Among these, clustering was not accounted for in 12 trials (10%). See Table 2.2.

Table 2.2: Analysis methods stratified by unit of analysis

	Cluster-level analysis	**Individual-level analysis**
Methods of analysis	6 (100)	117 (100)
Generalized estimating equations	-	27 (23.1)
Mixed effects models	-	55 (47.0)
Repeated measures analysis of variance	-	5 (4.3)
Generalized linear model with sandwich variance	-	16 (13.7)
Chi square accounting for clustering	-	1 (0.8)
Survival analysis accounting for clustering	-	1 (0.8)
Other methods ignoring clustering [a]	-	12 (10.3)
Weighted regression [b]	1 (16.7)	-
Other methods without weighting [a]	5 (83.3)	-
Methods of analysis when non-adherence was addressed	1 (100)	18 (100)
Generalized estimating equations	-	4 (22.2)
Mixed effects models	-	9 (50.0)
Generalized linear model with sandwich variance	-	4 (22.2)
t-test ignoring clustering [c]	-	1 (5.6)
Unweighted t-test [d]	1 (100)	-

The numbers in brackets are the column percentages. [a] Generalized linear model, analysis of variance, analysis of covariance, T-test, Mann-Whitney U test, Chi square test. [b] Number of events (cluster size) used as weights (Buttha et al [47]). The use of weigths is applicable when cluster-level summaries analysis is performed while accounting for clustering may be required for individual-level analysis. [c] t-test with multiple testing adjustment but ignoring clustering was applied to perform a per protocol analysis at individual-level (Neuzil et al. [48]). [d] Per protocol analysis with unweighted t-test comparing rates at cluster level (Tagbor et al. [49]).

2.3.2 The reporting and handling of non-adherence

Sixty-one reports (50%) included information on adherence: full adherence was reported in 5 trials while the remaining 56 trials reported some form of non-adherence to the allocated treatment. Table 2.3 reports the adherence characteristics of these trials. The reporting of adherence was more frequent in interventions of short duration (57%) compared to those of long duration (47%). Forty-four trials (72%) used a binary treatment adherence definition, with only one report justifying the threshold used for this dichotomisation. Five trials (8%) recorded non-adherence as a continuous variable, while the remaining 12 trials (20%) gave no details on the definition of adherence used. Only 11 trials (9%) provided a flow chart with complete information on how many clusters and/or individuals received the assigned treatment. Nine trials reporting non-adherence performed adverse events analysis.

Non-adherence at the cluster level was reported in 15 trials (24%), with a further 4 (6%) studies reporting full cluster adherence. Non-adherence at the individual level was acknowledged in 41 trials (71%) out of 58 trials that provided information on adherence (full adherence or presence of non-adherence), while one trial (2%) reported full adherence. No study reported the use of an intra-cluster correlation coefficient for adherence.

2.3.2.1 Adherence by allocated groups

Active group: Five studies provided the percentage of cluster-level non-adherence, with a median (1st–3rd quartiles) of 44.8% (33%–50%), with a further 10 reporting cluster non-adherence without further details. At the individual level, 30 (73%) out of 41 studies reported this, with a corresponding median (1st–3rd quartiles) of 15% (9%–24%).

Control group: Adherence to the control protocol was less frequently reported; 5 trials stated full adherence, while a further 15 studies reported some form of non-adherence. Cluster-level non-adherence was reported in one trial, while full adherence was reported in a further 4 trials. At the individual level, 19 trials reported control-treatment non-adherence, with full adherence in one study.

2.3.2.2 Adherence-adjusted analyses

Fifteen trials performed PP analyses, with the remaining 4 studies carrying out AT analyses either as primary or secondary analyses. No study reported LATE estimates. Amongst the 9 studies with a safety outcome, 4 trials performed a PP analysis [48, 50–52], with a further trial using an AT analysis [53]. Two studies did not account for clustering in their adherence-adjusted analyses [48, 49]. No study reported the assumptions necessary for their adherence-adjusted analyses to be unbiased causal treatment estimates. In any case, none of these studies was double-blinded. I summarise some of the characteristics of these adherence-adjusted analyses in Table 2.4.

Table 2.3: Reporting of non-adherence by length of intervention, randomised arm and level of adherence.

	One-off	Short term	Long term	Total
Reporting of any non-adherence, n (%)	**5 (100)**	**35 (100)**	**83 (100)**	**123 (100)**
Non-adherence reported in both active and control groups	-	4 (11.4)	16 (19.3)	20 (16.2)
Non-adherence reported in active group only	2 (40.0)	15 (42.9)	19 (22.9)	36 (29.3)
Non-adherence reported in control group only	-	-	-	-
Full adherence reported	-	1 (2.9)	4 (4.8)	5 (4.1)
Not reported	3 (60.0)	11 (31.4)	36 (43.4)	50 (40.6)
Unclear	-	4 (11.4)	8 (9.6)	12 (9.8)
Trials with adherence at cluster level	**2 (100)**	**21 (100)**	**42 (100)**	**65 (100)**
Non-adherence reported in both active and control groups	-	-	1 (2.4)	1 (1.5)
Non-adherence reported in active group only	-	9 (42.9)	5 (11.9)	14 (21.5)
Non-adherence reported in control group only	-	-	-	-
Full adherence reported	-	1 (4.7)	3 (7.1)	4 (6.2)
Not reported	2 (100)	9 (42.9)	29 (69.1)	40 (61.6)
Unclear	-	2 (9.5)	4 (9.5)	6 (9.2)
Trials with adherence at individual level	**3 (100)**	**14 (100)**	**41 (100)**	**58 (100)**
Non-adherence reported in both active and control groups	-	4 (28.6)	15 (36.6)	19 (32.8)
Non-adherence reported in active group only	2 (66.7)	6 (42.9)	14 (34.1)	22 (38.0)
Non-adherence reported in control group only	-	-	-	-
Full adherence reported	-	-	1 (2.4)	1 (1.7)
Not reported	1 (33.3)	2 (14.3)	7 (17.1)	10 (17.2)
Unclear	-	2 (14.3)	4 (9.8)	6 (10.3)
Percentage of non-adherence at cluster level[a]				
Total number of trials reporting non-adherence, n (%)	**-**	**9 (100)**	**6 (100)**	**15 (100)**
Trials reporting % of non-adherence in active group, n (%)	-	2 (22.2)	3 (50.0)	5 (33.3)
Median % of non-adherence in active group[b]	-	37.4 (30–44.8)	50 (33–80)	44.8 (33–50)

Continued on next page

Table 2.3 Continued

	One-off	Short term	Long term	Total
Percentage of non-adherence at individual level				
Total number of trials reporting non-adherence, n (%)	**2 (100)**	**10 (100)**	**29 (100)**	**41 (100)**
Trials reporting % of non-adherence in active group, n (%)	2 (100)	7 (70.0)	21 (72.4)	30 (73.2)
Median % of non-adherence in active group[b]	16.5 (0.5–32.4)	13.7 (5.3–25)	15 (10–20)	15 (9–24)
Total number of trials reporting non-adherence, n (%)	**2 (100)**	**10 (100)**	**29 (100)**	**41 (100)**
Trials reporting % of non-adherence in control group, n (%)	-	3 (30.0)	11 (37.9)	14 (34.1)
Median % of non-adherence in control group[b]	-	8.1 (1.7–32)	8.2 (3.4–20)	8.2 (3.4–20)
Total number of trials, n (%)	**5 (100)**	**35 (100)**	**83 (100)**	**123 (100)**
Flow chart with adherence information	1 (20.0)	4 (11.4)	6 (7.2)	11 (8.9)
Flow chart without adherence information	1 (20.0)	19 (54.3)	65 (78.3)	85 (69.1)
No flow char	3 (60.0)	12 (34.3)	12 (14.5)	27 (22.0)
Adherence type, n (%)[c]	**2 (100)**	**20 (100)**	**39 (100)**	**61 (100)**
Binary adherence	2 (100)	14 (70.0)	28 (71.8)	44 (72.1)
Continuous adherence	-	2 (10.0)	3 (7.7)	5 (8.2)
Unclear	-	4 (20.0)	8 (20.5)	12 (19.7)
Trials using adherence-adjusted methods, n (%)	**1 (100)**	**4 (100)**	**14 (100)**	**19 (100)**
Per protocol	1 (100)	4 (100)	10 (71.4)	15 (78.9)
As treated	-	-	4 (28.6)	4 (21.1)

[a] No report provided non-adherence % at cluster level in the control group. [b] Numbers in brackets are the 1st and 3rd quartiles. [c] Total number of trials reporting non-adherence or full adherence.

2.4 Summary of findings

This is the first systematic review of reporting practices of non-adherence with randomised treatment in CRTs. Our findings show that about half of the studies include information on treatment adherence, but details on numbers of clusters and individuals that adhered to the intended treatment are often incomplete. Schulz et al. [67, 68] found that trials reporting exclusions after treatment initiation (i.e. deviations from protocol) tend to be of higher methodological quality than those that did not report it. This is known as the "exclusion paradox". It is therefore possible that those studies that did not report on adherence also experienced protocol deviations. On this basis, I estimate that in this study the proportion of trials with non-adherence lies within the range 23% to 94% at the cluster level and 71% to 98% at the individual level. In addition, I found that studies tended to report more often adherence at the individual level. This potential under-reporting may be due to the complexity of defining adherence in CRTs, and that as CRTs are often pragmatic in nature, recording adherence to treatment protocol is not often a primary concern.

Amongst the studies reporting non-adherence, only one-third specified departures from protocol in the control group. This has to be interpreted in light of the fact that in our review, "usual care" was used as control in approximately three quarters of studies, and that defining and measuring adherence to "usual care" may be difficult or impractical. In general, the nature of the departures from protocol was very poorly reported, and it was not possible to ascertain whether alternative treatments to those in the trial, i.e. not originally included in the design of the study, were taken. Knowledge of the alternative regimes followed by those individuals who did not adhere to their allocated treatment is important if we want to judge the impact of such non-adherence on the causal interpretation of an ITT analysis. If no external treatments are available, then the ITT estimate will be diluted towards the null, when compared with the true treatment effect. Moreover, the reported non-adherence details (numbers initiating and completing the treatment protocol, period of discontinuation, etc.) were often inadequate for a meaningful interpretation of

Table 2.4: Details of the adherence-adjusted analyses performed

Study	Reason	Type	Differences in inference
Per protocol			
Acolet et al.[54]	Exploratory	Binary	PP not shown, stated similar to ITT
Auger et al. [55]	Unclear	Binary	ITT not done
Beer et al. [56]	Unclear	Binary	Evidence of effect with PP, but not with ITT
Bickman et al.[57]	Unclear	Binary	No change
Boorsma et al.[50][c]	Unclear	Binary	Evidence of effect with PP, but not with ITT
Cooke et al. [58]	Unclear	Binary[a]	ITT not done
Cutrer et al. [59]	Unclear	Binary	ITT not done
Dangour et al. [51][c]	Exploratory	Binary	No change
Estrada et al.[60]	Unclear	Binary	No change
Luoto et al.[52][c]	Unclear	Binary	No change
Neuzil et al.[48][c,d]	Safety	Binary	No change
Smith et al. [61]	Additional analyses	Binary	No change
Tagbor et al. [49][d]	Unclear	Binary[b,d]	Evidence of effect with PP, but not with ITT
Taveras et al. [62]	Unclear	Binary	No change
Zurovac et al.[63]	Unclear	Binary	No change
As-treated			
Stiell et al. [64]	Additional analyses	Binary	No change
Zamorano et al. [53][c]	Efficacy	Binary	ITT not done
LaBella et al. [65]	Unclear	Continuous	Evidence of effect with ITT, but not with AT
Levine et al. [66]	Unclear	Continuous	AT not shown

[a] The threshold chosen to define the binary non-adherence was based on a previous study. [b] All possible definitions of binary adherence explored (>1 dose,>2 doses and full exposure) [c] Carried out a safety outcome analysis. [d] Failed to adjust for clustering in the analysis.

the study results.

All of the studies reporting adherence-adjusted estimates performed PP or AT analyses, without discussing the assumptions necessary to result in unbiased causal estimates and their plausibility [18, 19]. No study performed a LATE estimation or any other statistical methods that are suitable under more plausible assumptions in the context of a randomised trial (such as assuming random allocation is a valid instrument) for unbiased causal estimation [24, 34].

2.5 Comparison with previous studies

A previous systematic review by Eldridge et al. [12] assessed the quality of design and reporting of CRTs published between 1997 and 2000 found that non-adherence was reported in about 24% of 152 studies. However, they did not investigate how non-adherence was addressed when estimating causal treatment effects. For individual randomised trials, Dodd et al. [11], in a review of 100 trials published in 2008 in five leading medical journals, found this percentage to be 98%. In contrast, the review performed by Zhang et al. [41], which considered individual randomised drug trials published in 2010, found a prevalence of non-adherence reporting of 46%. Both of these results are thus in line with the lower and upper bounds found in our study. These two previous individually-randomised trials reviews noted a lack of justification in the threshold used in defining a binary measure of non-adherence [11, 41]. In the present review, only one justified this choice.

The median percentage of individual-level non-adherence reported by the CRTs included here was 13%. Similar median percentages of non-adherence were found in previous reviews of adherence in individually randomised trials, 10–20% in Dodd [11], and 11.6% in Zhang [41]. While the latter reported finding a monotonic trend of adherence with regards to intervention duration [41], I did not find any such trend. This could be because adherence was not clearly reported in over 40% of both long and short-term interventions. In fact, in view of the "exclusion paradox", non-adherence with short-term interventions could be as high as the non-adherence reported in long-term interventions.

Only 3% of the studies included in the present review presented an adherence-adjusted analysis as primary, with the great majority reporting an ITT approach. Of those studies assessing treatment efficacy, PP analysis was the most used. Dodd et al. [11] also found that the majority of studies attempting to adjust for non-adherence in an analysis used PP.

Although the extended CONSORT statement for CRTs [14, 39] recommends reporting the numbers of clusters and individuals randomised and receiving their assigned treatment, I found that the reporting of these numbers was low (9%). This is in contrast to the results reported by Dodd et al. [11], who found that 58% stated the number of participants actually initiating their allocated treatment. A possible explanation may be the lower adherence to CONSORT guidelines among CRT reports [69] as well as the extra complexities of defining, measuring and recording adherence at both levels.

Chapter 3.

Introduction to cluster-level summary approaches in CRTs

3.1 Introduction

CRT investigators are sometimes interested in evaluating the effect of their intervention on the whole cluster, rather than the individual units. Such analyses may require a reduction of the hierarchical structure of clustered data to a single-level data. This is achievable by first constructing summary statistics of variables (for example, means or proportions where appropriate) for each cluster, and then performing statistical analyses on those summaries. Estimates obtained from such approaches are at the cluster level [8], with summaries taking values on a continuous scale. Implementation of cluster-level (CL) summary analyses may require accounting for the likely varying precision of CL summaries, which if substantial, would induce heteroscedasticity (non-constant residual variance). Standard statistical methods such as ordinary least squares (OLS) regression assume homoscedasticity (constant residual variance). Failure to address heteroscedasticity, if present, results in wrong inferences [70, 71]. Various re-weighting approaches such as cluster size (CS) and minimum-variance (MV) weights [72–74], and the use of Huber-White (so-called sandwich or robust) standard errors (SEs) are available to account for heteroscedasticity, if present. CL-summary analyses are known to be inefficient [75], but adequate weighting may improve efficiency [8, 76].

The current chapter introduces different approaches that have been proposed for CL-summary analyses in the context of ITT analysis [8]. Although the chapter is about CL-summary approaches, I present an overview of random effects linear modelling which uses the hierarchical structure of the data in Section 3.2 before moving to

CL-summary analyses. Then, section 3.3 introduces CL-summary approaches and section 3.4 presents a summary of the chapter.

I consider a two-arm CRT, with n individual units indexed by i, in J clusters indexed by j, each of size n_j. Let Z_j denote the binary treatment randomly allocated at the cluster level with probability 0.50. Let Y_{ij} represent the continuous or binary outcome, X_{ij} a set of explanatory variables measured for individual i in cluster j, and W_j a set of explanatory variables measured for cluster j.

3.2 Overview of random effects linear regression

Standard regressions assume that the observations are independent and identically distributed. The independence assumption is violated when data are clustered *i.e.* observations are on grouped individuals or are repeated measurements on the same individuals. Ignoring the clustering leads to underestimating the SEs for cluster-level covariates which include the treatment allocation variable, and therefore to incorrect inferences. Random effects modelling adequately handles clustered data (when correctly specified) and is applicable to any type of dependent variables (continuous, binary, counts, etc.). Here, we focus on linear regression models with random intercepts. These models are not part of CL-summary approaches but are first introduced in order to clarify notions such as the intracluster correlation coefficient and the between-cluster variance used later for CL-summary analyses.

Random intercept linear models relax the standard assumption of independence across individual units by allowing for cluster-specific intercepts. Here, Y_{ij} represents only a continuous outcome. A random-intercept linear model for ITT analysis, with covariate adjustment for generality, is as follows [8, 77, 78]

$$Y_{ij} = \gamma_0 + \gamma_{\mathrm{z}} Z_j + \gamma_{\mathrm{w}} W_j + \gamma_{\mathrm{x}} X_{ij} + \upsilon_j + \epsilon_{ij} \tag{3.1}$$

where γ_0 is the mean intercept shared by all clusters and individual units when Z, W and X are equal to 0; γ_{z} is the IL-ITT effect, conditional on W and X. γ_{w} and γ_{x} are the change in mean outcome induced by a unit increase or a change in level of W_j and X_{ij}, but are of little interest in randomised trials. υ_j is the random intercept

(or level-2 residual) for cluster j and ϵ_{ij} the IL (level-1) residuals. It is assumed that $\upsilon_j \sim N(0, \sigma_\upsilon^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and $\text{Cov}(\epsilon_{ij}, \upsilon_j | X_{ij}, W_j, Z_j) = 0$. More explicitly, $\mathbb{E}(\epsilon_{ij} | X_{ij}, W_j, Z_j, \upsilon_j) = 0$ and $\mathbb{E}(\upsilon_j | X_{ij}, W_j, Z_j) = 0$. This implies that the level-1 residual ϵ_{ij} and random intercept υ_j are both uncorrelated with the explanatory variables. It is also assumed that $\text{Cov}(\epsilon_{ij}, \epsilon_{i'j'} | X_{ij}, W_j, Z_j, \upsilon_j, X_{i'j'}, W_{j'}, Z_{j'}, \upsilon_{j'}) = 0$ for any $i \neq i'$ or $j \neq j'$ and $\text{Cov}(\upsilon_j, \epsilon_{ij'} | X_{ij}, W_j, Z_j, X_{i'j'}, W_{j'}, Z_{j'}) = 0$ for any $j \neq j'$. Note that it is not necessary to condition on Z_j because randomisation ensures that Z_j is independent of υ_j and ϵ_{ij}. Covariates adjustment in ITT analysis is not to control for confounding, but rather for improving precision.

The regression coefficients are consistent if the mean structure is correctly specified *i.e.* correct functional form and covariates, and $\text{Cov}(\epsilon_{ij}, \upsilon_j) = 0$. The additional assumption that the covariance structure (variances and covariances) of the total residual $(\upsilon_j + \epsilon_{ij})$ is correctly specified, is necessary to ensure the consistency of the model-based SEs. When both mean and covariance structures are correctly specified, the coefficients estimation is efficient. The coefficients are unbiased if the mean structure is correct and the total residuals have a symmetrical distribution [78]. Estimations are often done by maximum likelihood (ML) or restricted (or residual) maximum likelihood (REML) [79]. ML estimation consists of maximizing the joint likelihood function of the coefficients and variance component [77]. REML unlike ML estimates the random-intercept variance accounting for the loss of degrees of freedom resulting from the estimation of the coefficients. REML provides unbiased estimates for the variance component whereas ML leads to a downward biased variance component when data are balanced, that is all clusters have the same size. However for unbalanced data, both REML and ML are biased. It is unclear which of REML and ML has the smaller mean square error, whether data are balanced or unbalanced [77, 78].

Note that the total residual variance (denoted σ^2) is $\sigma^2 = \sigma_\upsilon^2 + \sigma_\epsilon^2$, where $\text{Var}(Y_{ij}) = \sigma^2$. The marginal intracluster correlation coefficient (ICC) for Y_{ij}, denoted ρ_y, is obtained from fitting equation (3.1) without regressors (Z_j included, *i.e.* an empty

model) and is given by $\rho_y = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2}$. Thus, the between-cluster variance of Y_{ij} can be expressed as $\sigma_v^2 = \rho_y \sigma^2$ and the within-cluster variance of Y_{ij} as $\sigma_\epsilon^2 = (1 - \rho_y)\sigma^2$.

3.3 ITT analysis on CL summaries

Another strategy to handle CRT data is to first compute CL-summary statistics and use these summaries as variables to analyse or include in a regression model for the estimation of CL-ITT. This approach is suitable when the treatment effect at the cluster level is of interest. I introduce two CL-summary approaches for CRTs, the unadjusted and adjusted CL-summary outcomes. ITT analysis with unadjusted or adjusted CL-summary outcome may require SEs that are robust to heteroscedasticity to ensure valid inferences. This can be achieved via adequate weighting strategy (*CS* or *MV* weights) or HW SEs. In this section, we formally present the ITT analysis using the unadjusted and adjusted CL-summary outcomes, combined with weighting and HW SEs.

I denote $\overline{Y}_j$ and $\overline{X}_j$ be the sample means of Y_{ij} and X_{ij} for cluster j respectively, *i.e.* $\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$ and $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$. If X_{ij} includes categorical variables, then for each of these variables, dummy variables will be created for each category except for one, and sample proportions generated for each of these dummies.

3.3.1 Unadjusted CL summaries

The unadjusted CL (*unCL*)-summary approach consists of generating CL-summary statistics, such as means or proportions where appropriate, of the observed outcome. These CL summaries are used in subsequent analyses. CL-summary statistics are continuous regardless of whether the original variable was binary or not. For categorical regressors with q levels, $(q-1)$ dummy variables can be created and proportions generated for each of these dummies. Thus, statistical methods such as t-test or *Wilcoxon-Mann-Whitney* test can be used to assess the ITT effect by comparing the CL summaries between the control and active treatment groups, if the trial investigators do not plan to adjust for covariates. A linear regression model can be fitted to adjust for covariates, but only CL regressors should be directly included *i.e.* W_j but not the mean of X_{ij}. It has been shown that including CL summaries

from IL variables as regressors is equivalent to using those regressors as instrumental variables and is therefore inappropriate [76].

Here, we consider CL-sample means or proportions, $\overline{Y}_j$ and $\overline{X}_j$. Z_j and W_j are measured at the cluster level. In fact, equation 3.1 can be written as $Y_{ij} = \mu_j + \gamma_{\mathrm{x}} X_{ij} + \epsilon_{ij}$ where $\mu_j = \gamma_0 + \gamma_{\mathrm{z}} Z_j + \gamma_{\mathrm{w}} W_j + \upsilon_j$ are the cluster means adjusted for differences in the individual-level covariate X_{ij} [76]. Likewise, $X_{ij} = \phi_0 + \phi_{\overline{\mathrm{x}}} \bar{X}_j + \psi_{ij} = \phi_0 + \overline{X}_j + \psi_{ij}$ as $\phi_{\overline{\mathrm{x}}}$ is always equal to 1 (where ψ_{ij} is a random error term with mean 0, assumed to be independently and identically distributed). Therefore, we have the following simultaneous equations to solve [76]:

$$\begin{aligned} Y_{ij} &= \gamma_0 + \gamma_{\mathrm{z}} Z_j + \gamma_{\mathrm{w}} W_j + \gamma_{\mathrm{x}} X_{ij} + \upsilon_j + \epsilon_{ij} \\ X_{ij} &= \phi_0 + \overline{X}_j + \psi_{ij} \end{aligned} \tag{3.2}$$

Equation 3.2 corresponds to an IV setting where $\overline{X}_j$ is an instrument for X_{ij}. The variability in Y_{ij} at the cluster level (adjusted for X_{ij}) is captured by the μ_j's. Regressing the estimated cluster means $\widehat{\mu}_j$ on the cluster-level covariates Z_j and W_j enables us to estimate the target treatment effect γ_{z} [76]. If we substitute X_{ij} by $\overline{X}_j$ and express Y_{ij} as functions of Z_j, W_j and $\overline{X}_j$, then we have the cluster means $\mu_j^\star = \gamma_0 + \gamma_{\mathrm{z}} Z_j + \gamma_{\mathrm{w}} W_j + \beta_{\overline{\mathrm{x}}} \overline{X}_j + \upsilon_j$ that are different from μ_j. The $\mu_j^\star$'s are not adjusted for X_{ij} and therefore using $\mu_j^\star$ as cluster means to explain the variability in Y_{ij} at the cluster level identifies a parameter that is different from the target treatment effect γ_{z} in equation 3.1 [76]. This is the rationale for using the adjusted CL-summaries covered in section 3.3.2, if we want to adjust the cluster means for adjusting individual-level covariates.

Since CL-summary outcomes are continuous, they can be assumed to be approximately normally distributed, especially when n_j is sufficiently large. The OLS regression model for estimating the ITT effect with CL-summary outcome $\overline{Y}_j$ as the dependent variable and with covariate adjustment is as follows

$$\overline{Y}_j = \alpha_0 + \alpha_{\mathrm{z}} Z_j + \alpha_{\mathrm{w}} W_j + \eta_j \tag{3.3}$$

where $j \in \{1, \ldots, J\}$, η_j is a random error term, assumed to be i.i.d., with mean 0

and constant variance σ_η^2. Note that X_{ij} is not included in equation (3.3). The ITT effect is α_Z. The OLS estimator $\widehat{\alpha}_k$ of the parameter α_k (here $\alpha_k = \alpha_0, \alpha_{\mathrm{z}}$ or α_{w}) is consistent *i.e.* $\widehat{\alpha}_k \to \alpha_k$ as the number of clusters n goes to infinity. OLS estimator is unbiased regardless the sample size if $\mathrm{Cov}(Z_j, \eta_j) = 0$ and $\mathrm{Cov}(W_j, \eta_j) = 0$ [80].

Equation (3.3) is equivalent to equation (3.1) without IL regressor X_{ij} as long as the functional form is linear. In such a case, by using the expectation of Y_{ij}, we can note that α_{z} and γ_{z} are equivalent and therefore can be interpreted as either CL-ITT or IL-ITT effects.

Let $\mathbf{w}_j$ be the vector of all CL regressors in equation (3.3) for cluster j including a vector column of 1 for intercept, *i.e.* $\mathbf{w}_j = \begin{bmatrix} 1 & Z_j & W_j \end{bmatrix}$ and $\mathbf{w'}_j$ its transpose, $\mathbf{w}_j$ has dimension $1 \times p$, where p is the number of parameters in equation (3.3). The vector of OLS estimator $\widehat{\alpha}_{\mathrm{ols}} = \begin{bmatrix} \widehat{\alpha}_0 & \widehat{\alpha}_Z & \widehat{\alpha}_W \end{bmatrix}'$ is given by

$$\widehat{\alpha}_{\mathrm{ols}} = \Big(\sum_{j=1}^{J} \mathbf{w'}_j \mathbf{w}_j \Big)^{-1} \Big(\sum_{j=1}^{J} \mathbf{w'}_j \overline{Y}_j \Big) \tag{3.4}$$

The asymptotic variance matrix of OLS estimator vector $\widehat{\alpha}_{\mathrm{ols}}$ is

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathrm{ols}}) = \Big(\sum_{j=1}^{J} \mathbf{w'}_j \mathbf{w}_j \Big)^{-1} \sum_{j=1}^{J} \widehat{\sigma}^2_{\eta_j} \mathbf{w'}_j \mathbf{w}_j \Big(\sum_{j=1}^{J} \mathbf{w'}_j \mathbf{w}_j \Big)^{-1} \tag{3.5}$$

where $\widehat{\sigma}^2_{\eta_j}$ is the estimated variance of the residual η_j from equation (3.3). Under independence and homoscedasticity assumptions, $\widehat{\sigma}^2_{\eta_j} = \widehat{\sigma}^2_{\eta}$ for any cluster j. Thus, equation (3.5) is simplified to equation (3.6) below where $\widehat{\sigma}^2_{\eta} = \frac{1}{J-p} \sum_{j=1}^{J} \widehat{\eta}^2$ [80].

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathrm{ols}}) = \widehat{\sigma}^2_{\eta} \Big(\sum_{j=1}^{J} \mathbf{w'}_j \mathbf{w}_j \Big)^{-1} \tag{3.6}$$

If the ITT analysis is unadjusted *i.e.* W_j not included in equation (3.3), then

$$\widehat{\alpha}_{\mathrm{z}} = \frac{\mathrm{Cov}(\overline{Y}_j, Z_j)}{\mathrm{Var}(Z_j)} \tag{3.7}$$

If W_j represents a single CL covariate in equation (3.3), then the expanded formula

of the ITT effect estimate is $\widehat{\alpha}_Z = \dfrac{\left(\sum_{j=1}^{J} W_j^2\right)\left(\sum_{j=1}^{J} Z_j \overline{Y}_j\right) - \left(\sum_{j=1}^{J} Z_j W_j\right)\left(\sum_{j=1}^{J} W_j \overline{Y}_j\right)}{\left(\sum_{j=1}^{J} Z_j^2\right)\left(\sum_{j=1}^{J} W_j^2\right) - \left(\sum_{j=1}^{J} Z_j W_j\right)^2}$.

As noted, the formulae in equations (3.4) and (3.6), well known by econometricians, are not that intuitive particularly in the presence of covariate adjustment. In fact, in multivariable linear regression, OLS coefficient for the k^{th} regressor is simply the ratio of: (i) the covariance of the outcome variable and the residuals from regressing that k^{th} regressor on all remaining covariates, to (ii) the variance of those residuals [76, 81, 82]. This is so-called "regression anatomy" [76, 83] and we can write the covariate-adjusted ITT effect estimate $\widehat{\alpha}_Z$ as

$$\widehat{\alpha}_{\mathrm{z}} = \frac{\mathrm{Cov}(\overline{Y}_j, \epsilon_{z_j})}{\mathrm{Var}(\epsilon_{z_j})} \tag{3.8}$$

where ϵ_{z_j} are the residuals from OLS regression of Z_j on all other CL covariates W_j included in equation (3.3) *i.e.* the residuals from the OLS model $Z_j = \delta_0 + \delta_{\mathrm{w}} W_j + \epsilon_{z_j}$. Intuitively, closely looking at equations (3.7) and (3.8), when the covariate-adjusted ITT effect is of interest, we first obtain a "transformed" Z_j (namely, the residuals ϵ_{z_j}) that is independent of all the CL covariates and then use this "transformed" Z_j as the single exposure or regressor for the ITT analysis. The proof of equation (3.8) is provided in appendix A.6 [76, 83].

As Z_j and W_j are independent by randomisation, then $\delta_{\mathrm{w}} = 0$ and the predicted value of Z_j is $\widehat{Z}_j = \mathbb{E}(Z_j)$. Thus, ϵ_{z_j} is the mean-centred Z_j *i.e.* $\epsilon_{z_j} = Z_j - \mathbb{E}(Z_j)$. Therefore, $\mathrm{Cov}(Y_j, \epsilon_{z_j}) = \mathrm{Cov}(Y_j, Z_j - \mathbb{E}(Z_j)) = \mathrm{Cov}(Y_j, Z_j)$. Hence, the covariate-adjusted and unadjusted ITT effect estimates are equivalent, supporting the claim that covariate-adjustment in the presence of randomised treatment is not to address confounding. Nevertheless, the covariate-adjusted and unadjusted ITT effect estimates may differ due to possible covariates imbalance in finite samples.

Generally, the variance of $\widehat{\alpha}_{\mathrm{z}}$ can also be written as [84]

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathrm{z}}) = \widehat{\sigma}_{\eta}^2 \Big/ \left((1 - R^2_{\mathrm{z|w}}) \sum_{j=1}^{J} (Z_j - \overline{Z})^2\right) \tag{3.9}$$

where $R^2_{\mathrm{z|w}}$ is the coefficient of determination (often denoted R^2) of Z_j regressed on

all CL covariates W_j included in equation (3.3). From equation (3.9), it appears clearly that – (i) the smaller the residual variance $\widehat{\sigma}^2_\eta$, the more efficient is the covariate-adjusted ITT effect estimate $\widehat{\alpha}_{\text{Z}}$, – (ii) the lower $R^2_{\text{Z|W}}$, the more efficient is $\widehat{\alpha}_{\text{Z}}$. By design, $Z_j \perp\!\!\!\perp W_j$ and thus, $R^2_{\text{Z|W}}$ is expected to be equal to 0 (smallest value). In CRTs with approximately equal number of clusters per group, $\overline{Z} \approx \frac{1}{2}$ and thus $\sum_{j=1}^{J}(Z_j - \overline{Z})^2 \approx \frac{J}{4}$. Therefore, $\widehat{\text{Var}}(\widehat{\alpha}_Z)$ can be approximated as follows

$$\widehat{\text{Var}}(\widehat{\alpha}_Z) \approx \frac{4\widehat{\sigma}^2_\eta}{J} \tag{3.10}$$

Increasing the number of clusters would improve the efficiency of the CL-ITT effect estimator $\widehat{\alpha}_{\text{Z}}$ [85]. Adjusting for CL covariates that are associated with the *unCL*-summary outcome $\overline{Y}_j$ would substantially reduce the residual variance $\widehat{\sigma}^2_\eta$ and subsequently lead to efficiency gain. Otherwise, when CL covariates are not associated with the outcome $\overline{Y}_j$, the loss of degrees of freedom coupled with the negligible reduction in the residual sum of squares undermines the efficiency, particularly when J is small [85, 86].

For hypothesis testing, $\frac{\widehat{\alpha}_{\text{Z}}}{[\widehat{\text{Var}}(\widehat{\alpha}_{\text{Z}})]^{\frac{1}{2}}} \sim t(J-p)$ if η_j is normally or at least asymptotically distributed. The efficiency of OLS estimator relies on the homoscedasticity and i.i.d. assumptions for the error term η_j. The lack of normality of the error term η_j is neither a threat for consistency nor efficiency of OLS estimators. However, when the normality assumption is satisfied, hypothesis testing and confidence intervals are reliable.

3.3.2 Adjusted CL summaries

As mentioned in the previous section, a drawback of regression analyses on CL-summary outcome is its inability to adequately adjust for IL regressors [76]. However, where there is interest in adjusting for covariates at the individual level when analysing CL summaries, a two-step procedure so-called "residual index" [87] or "covariate-adjusted residual" [8] can be applied. First, an IL regression analysis of the outcome, while ignoring clustering, is performed incorporating all the relevant covariates into the regression model except the treatment variable. Then, outcome

residuals are predicted and used to compute CL-summary statistics which serve as dependent variable in subsequent analyses. I refer to this as adjusted CL (*adCL*)-summary outcome. The *adCL*-summary approach varies according to the type of outcome variable. I present this for continuous (via OLS regression) and binary (via logistic regression) outcomes.

3.3.2.1 Continuous outcome

The *adCL*-summary approach for continuous outcome Y_{ij} is as follows

- In the first step, OLS regression on Y_{ij} is performed as

$$Y_{ij} = \gamma_{1_0} + \gamma_{1_W} W_j + \gamma_{1_X} X_{ij} + e_{1_{ij}} \quad \text{with} \quad e_{1_{ij}} \sim N(0, \sigma^2_{e_1}) \tag{3.11}$$

  and the outcome residuals are obtained as follows

$$\widehat{e}_{1_{ij}} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - (\widehat{\gamma}_{1_0} + \widehat{\gamma}_{1_W} W_j + \widehat{\gamma}_{1_X} X_{ij}) \tag{3.12}$$

  where $\widehat{Y}_{ij} = (\widehat{\gamma}_{1_0} + \widehat{\gamma}_{1_W} W_j + \widehat{\gamma}_{1_X} X_{ij})$ is the predicted outcome at the individual level. The covariate-adjusted residuals $\widehat{e}_{1_{ij}}$ would be similar on average in the control and active groups under the null hypothesis of no treatment effect [8].

- In the second step, CL-summary statistic of outcome residuals $\widehat{e}_{1_{ij}}$, here means, are generated as $\overline{\widehat{e}}_j = \frac{1}{n_j} \sum_{j=1}^{n_j} \widehat{e}_{1_{ij}}$. These CL summaries are then used as dependent variables in a OLS regression for instance, as follows

$$\overline{\widehat{e}}_j = \alpha_0 + \alpha_{\text{Z}} Z_j + \eta_j \tag{3.13}$$

where $\eta_j \sim N(0, \sigma^2_\eta)$.

Caution is needed when drawing inference from equation (3.13). The degrees of freedom is reduced by the number of CL regressors W_j included in equation (3.11) and the t- statistic related to $\hat{\alpha}_{\text{Z}}$ must be adjusted accordingly [8]. In the absence of CL covariates in equation (3.11), the degrees of freedom related to $\hat{\alpha}_{\text{Z}}$ are $df = n-2$. However, when CL covariates are included in equation (3.11), then $df = n - 2 - k$, where k is the number of CL covariates in equation (3.11). Existing statistical software do not automatically correct the degrees of freedom and therefore, it is the analysts' responsibility to do so. Alternatively, we can fit equation (3.11) without

CL covariates W_j and later on, include W_j as regressor in equation (3.13) as follows [8].

$$Y_{ij} = \gamma_{1_0} + \gamma_{1_X} X_{ij} + e_{1_{ij}} \tag{3.14}$$

$$\bar{\hat{e}}_j = \alpha_0 + \alpha_{\mathrm{Z}} Z_j + \alpha_W W_j + \eta_j \tag{3.15}$$

The *adCL*-summary ITT approach based on equations (3.14) and (3.15) do not require the analyst to adjust the degrees of freedom. Results from standard software can then directly be reported.

3.3.2.2 Binary outcome

Consider Y_{ij} to be binary. The two steps for *adCL*-summary approach are summarised as follows [8].

- In the first step, a standard logistic regression is fitted as shown below

$$\begin{aligned} Y_{ij} \sim \mathrm{Bern}(\pi_{ij}) \quad \text{with} \quad \pi_{ij} = P(Y_{ij} = 1) \\ \mathrm{logit}(\pi_{ij}) = \gamma_0 + \gamma_{\mathrm{W}} W_j + \gamma_{\mathrm{X}} X_{ij} \end{aligned} \tag{3.16}$$

  and then, the observed (O_j) and expected (E_j) number of events are obtained for each cluster as follows

$$O_j = \sum_{i=1}^{n_j} Y_{ij} \quad \text{and} \quad E_j = \sum_{i=1}^{n_j} \hat{\pi}_{ij} = \sum_{i=1}^{n_j} \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_{\mathrm{W}} W_j + \hat{\gamma}_{\mathrm{X}} X_{ij})}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_{\mathrm{W}} W_j + \hat{\gamma}_{\mathrm{X}} X_{ij})} \tag{3.17}$$

  where $\hat{\pi}_{ij}$ is the predicted probability that $Y_{ij} = 1$. $\hat{\gamma}_0$, $\hat{\gamma}_{\mathrm{W}}$ and $\hat{\gamma}_{\mathrm{X}}$ are the estimated coefficients from equation (3.16).

- In the second step,

  if the risk ratio is of interest, then the *adCL*-summary outcome is

$$\bar{\hat{e}}_j = \frac{O_j}{E_j} \tag{3.18}$$

  if the risk difference is of interest, then the *adCL*-summary outcome is

$$\bar{\hat{e}}_j = \frac{O_j - E_j}{n_j} \tag{3.19}$$

The ITT analysis can be carried out using the following OLS regression.

$$\overline{\widehat{e}}_j = \alpha_0 + \alpha_{\mathrm{z}} Z_j + \eta_j \tag{3.20}$$

Like in section 3.3.2.1, for convenience regarding the degrees of freedom adjustment due to the inclusion of CL covariates, we can replace equation (3.16) with (3.21), and equation (3.20) with (3.22) below.

$$\begin{aligned} Y_{ij} \sim Bern(\pi_{ij}) \quad \text{with} \quad \pi_{ij} = P(Y_{ij} = 1) \\ \text{logit}(\pi_{ij}) = \gamma_0 + \gamma_{\mathrm{x}} X_{ij} \end{aligned} \tag{3.21}$$

and,

$$\overline{\widehat{e}}_j = \alpha_0 + \alpha_{\mathrm{z}} Z_j + \alpha_{\mathrm{w}} W_j + \eta_j \tag{3.22}$$

α_{z} represents the adjusted CL-risk difference ITT effect if the *adCL*-summary outcome is generated via equation (3.19) and the adjusted CL-risk ratio ITT effect if the *adCL*-summary outcome is computed using equation (3.18).

Note that, irrespective of the CL-summary approach adopted, the OLS estimation procedure of $\widehat{\alpha}_{\mathrm{z}}$ as presented in section 3.3.1 remains applicable. OLS assumptions may not be met when CL summaries are analysed. I present various ways of handling inferential flaws in the next section, when there are departures from key OLS assumptions. These include the use of weighting and/or heteroscedastic-robust standard errors.

3.3.3 Obtaining valid inferences

CL-summary analyses may lead to invalid inferences because of heteroscedasticity due to the varying cluster size. Violation of the key assumption of homoscedasticity also jeopardizes the efficiency of OLS estimator. CL-summary analyses, compared to their IL counterparts, are known to be inefficient [75]. Estimation by weighted least squares (WLS), where the weights are defined either by the *CS* or by the *MV* weights can improve efficiency [76]. The use of weights and/or heteroscedasticity-robust SEs helps in tackling heteroscedasticity. I present here different weighting strategies and the popular HW SEs estimator [70].

3.3.3.1 Heteroscedasticity-robust standard errors

The HW SE estimator is consistent even in the presence of substantial heteroscedasticity and is recommended [70,80,88]. HW SEs are derived from equation (3.5) where $\widehat{\sigma}^2_{\eta_j}$ is allowed to be different across clusters [70], as follows

$$\widehat{\mathrm{Var}}_{\mathrm{HW}}(\widehat{\alpha}_{\mathrm{ols}}) = \Big(\sum_{j=1}^{J} \mathbf{w_j'w_j}\Big)^{-1} \sum_{j=1}^{J} \widehat{\eta}^2_j \mathbf{w_j'w_j} \Big(\sum_{j=1}^{J} \mathbf{w_j'w_j}\Big)^{-1} \tag{3.23}$$

HW SEs are consistent even if $\widehat{\eta}^2_j$ is not a consistent estimator for $\sigma^2_{\eta_j}$ [70, 80]. However, HW SEs perform poorly in small samples of clusters; they are biased downwards and provide a coverage substantially below the nominal rate [71].

3.3.3.2 Weighting strategies

Consider equation (3.1) without regressors as follows

$$Y_{ij} = \gamma_0 + \upsilon_j + \epsilon_{ij} \tag{3.24}$$

We can write the CL-sample means outcome as $\overline{Y}_j = \gamma_0 + \upsilon_j + \frac{1}{n_j}\sum_{i=1}^{n_j} \epsilon_{ij}$. Thus, the variance of $\overline{Y}_j$ is $\mathrm{Var}(\overline{Y}_j) = \mathrm{Var}(\upsilon_j) + \frac{1}{n_j^2}\sum_{i=1}^{n_j} \mathrm{Var}(\epsilon_{ij}) = \sigma^2_\upsilon + \frac{\sigma^2_\epsilon}{n_j}$ [89]. Using the expression of σ^2_υ and σ^2_ϵ as functions of ρ_y and the total variance of Y_{ij} (*i.e.* σ^2) presented in section 3.2, we can re-write $\mathrm{Var}(\overline{Y}_j)$ as follows.

$$\mathrm{Var}(\overline{Y}_j) = \frac{1 + \rho_y(n_j - 1)}{n_j}\sigma^2 \tag{3.25}$$

Let ω_j be the weight assigned to cluster j. *CS* weighting, sometimes referred to as equal weights to individual units [73], consists of assigning the cluster size n_j as the weight to cluster j, *i.e.* $\omega_j = n_j$. *MV* weights are proportional to the inverse of the variance of CL-summary outcome and expressed as $\omega_j = \dfrac{n_j}{1 + \rho_y(n_j - 1)}$ [73,74]. *MV* weights minimize the variance of weighted estimators [90]. *MV* weights approximate *CS* weights when $\rho_y \approx 0$, whereas *MV* weights are approximately equivalent to *no* weights *i.e.* "equal weights to clusters" if $\rho_y \approx 1$ [89]. These equivalences can have practical implications when the variance of υ_j cannot be consistently estimated, for example when the number of clusters is small. In such setting, *CS* and *no* weights are viable alternatives. When clusters are large, weighting by cluster size is

inefficient [74]. When ρ_y is large, say 0.6 or greater [91], "equal weights to clusters" are preferable over *CS* weights as there is little variability within clusters [92].

3.3.3.3 Weighted least squares estimation

WLS is a special case of generalised least squares (GLS) estimation [93] where the off-diagonal elements of the residuals covariance matrix are 0, *i.e.* residuals are assumed to be independent but are allowed to be heteroscedastic. The WLS estimation of ITT effect is equivalent to using OLS estimation on the transformed equation below, where the weight ω_j are assumed to be known.

$$\overline{Y}_j\sqrt{\omega_j} = \alpha_0\sqrt{\omega_j} + \alpha_Z Z_j\sqrt{\omega_j} + \alpha_W W_j\sqrt{\omega_j} + \widetilde{\eta}_j \tag{3.26}$$

where $\widetilde{\eta}_j = \eta_j\sqrt{\omega_j}$ is the error term. As noted in equation (3.26), WLS estimation consists of fitting OLS where the outcome and covariates are rescaled by the square root of the weight ω_j, the intercept constrained to be 0 and the square root of the weights also included as covariate.

Let us denote $\overline{Y}_j\sqrt{\omega_j}$ by $\widetilde{\overline{Y}}_j$, $Z_j\sqrt{\omega_j}$ by $\widetilde{Z}_j$ and $W_j\sqrt{\omega_j}$ by $\widetilde{W}_j$. Let Ω_j be a variable whose values are $\sqrt{\omega_j}$. The $p\times 1$ vector of CL regressors $\mathbf{w_j} = \begin{bmatrix}1 & Z_j & W_j\end{bmatrix}$ introduced in section 3.3.1 becomes $\widetilde{\mathbf{w}}_j = \begin{bmatrix}\Omega_j & \widetilde{Z}_j & \widetilde{W}_j\end{bmatrix}$. Like the OLS setting (section 3.3.1), the unadjusted (*i.e.* without CL covariates adjustment) WLS ITT effect estimate can be obtained as

$$\widehat{\alpha}_Z = \frac{\mathrm{Cov}(\widetilde{\overline{Y}}_j, \widetilde{Z}_j)}{\mathrm{Var}(\widetilde{Z}_j)} \tag{3.27}$$

Using the "regression anatomy" formula, we can write the adjusted WLS ITT effect estimate $\widehat{\alpha}_Z$ as

$$\widehat{\alpha}_Z = \frac{\mathrm{Cov}(\widetilde{\overline{Y}}_j, \epsilon_{\widetilde{z}_j})}{\mathrm{Var}(\epsilon_{\widetilde{z}_j})} \tag{3.28}$$

where $\epsilon_{\widetilde{z}_j}$ are the residuals from OLS regression of $\widetilde{Z}_j$ on all other rescaled CL covariates $\widetilde{W}_j$ and Ω_j, but without intercept *i.e.* from the OLS model $\epsilon_{\widetilde{z}_j} = \delta_\omega\Omega_j + \delta_{\widetilde{W}}\widetilde{W}_j + \epsilon_{\widetilde{z}_j}$. Proof of equations (3.27) and (3.28) are shown in appendix A.7 [83].

The variance of the error terms $\widetilde{\eta}_j$ is estimated like in OLS regression as

$$\widehat{\mathrm{Var}}(\widetilde{\eta}_j) = \widehat{\sigma}^2_{\widetilde{\eta}} = \frac{1}{J-p}\sum_{j=1}^{J}\left(\overline{Y}_j\sqrt{\omega_j} - \widehat{\alpha}_0\sqrt{\omega_j} - \widehat{\alpha}_Z Z_j\sqrt{\omega_j} - \widehat{\alpha}_W W_j\sqrt{\omega_j}\right)^2 \tag{3.29}$$

The asymptotic variance matrix of WLS estimator vector $\hat{\alpha}_{\text{wls}} = \begin{bmatrix} \widehat{\alpha}_0 & \widehat{\alpha}_Z & \widehat{\alpha}_W \end{bmatrix}'$ is estimated as [88]

$$\widehat{\text{Var}}(\hat{\alpha}_{\text{wls}}) = \widehat{\sigma}^2_{\widetilde{\eta}} \Big(\sum_{j=1}^{J} \widetilde{\mathbf{w}}'_{\mathbf{j}} \widetilde{\mathbf{w}}_j \Big)^{-1} \tag{3.30}$$

Asymptotically, $\dfrac{\widehat{\alpha}_Z}{[\widehat{\text{Var}}(\widehat{\alpha}_Z)]^{\frac{1}{2}}} \sim t(J-p)$, where $\widehat{\text{Var}}(\widehat{\alpha}_Z)$ is the estimated WLS variance of $\widehat{\alpha}_Z$. If the homoscedasticity assumption conditional on the CL regressors is satisfied, OLS estimation is generally more efficient than WLS in finite samples. However, in the presence of substantial conditional heteroscedasticity, WLS is more efficient than OLS [88]. Testing for conditional heteroscedasticity, using for example the test of Breusch and Pagan [94] or White [70], may guide the choice between OLS or WLS estimation. OLS estimation would be preferred if there is no evidence of heteroscedasticity [88]. Regardless of the estimation method, OLS or WLS, it is recommended to also use HW SEs [70, 88].

3.4 Summary

The present chapter is an introduction of ITT analyses, based on individual-level data (random intercept models) or CL-summary data. CL-summary ITT analyses are known to perform well under various settings [8], but the performance of CL-TSLS with alternative weighting while adjusting for cluster-level and/or individual level covariates, has not been explored. Two CL-summary approaches based on unadjusted and adjusted CL-summary outcome to estimate the ITT effect of a randomised treatment are introduced. The former is simple whereas the latter offers the possibility of adequately adjusting for both CL and IL covariates as often desired by trial investigators to increase efficiency. It is possible to adjust for CL covariates before generating the CL-summary outcome. However, this requires an adjustment of the degrees of freedom when performing the TSLS estimation. An alternative way to avoid correcting the degrees of freedom oneself is to only adjust for IL covariates before computing the CL-summary outcome.

Chapter 4.

Estimation of local average treatment effect at the cluster level in CRTs

4.1 Introduction

Chapter 3 introduced CL-summary approaches in the context of ITT analysis. However, analysis at the cluster level may extend to the causal treatment effect when non-adherence is present [13, 36, 95]. In such a setting, two-stage least squares (TSLS) can be used to estimate the local average treatment effect (LATE), that is the causal effect of actually receiving the treatment [27]. While a CL method based on the Wald estimator [31] has previously been proposed [13], the performance of CL-TSLS to estimate LATE with alternative weighting while adjusting for CL and/or IL covariates, has not been explored.

I present here TSLS estimation and the Wald estimator of CL-LATE using CL-summary techniques introduced earlier in Chapter 3. The Wald estimator and TSLS are originally developed in the econometric literature, but little is known about their performance using CL summaries. This chapter is organized as follows. Section 4.2 outlines the required assumptions for the identification of CL-LATE. Section 4.3 presents TSLS estimation of CL-LATE. Section 4.4 introduces the Wald estimator with the Schochet-Chiang approach for estimating CL-LATE and its standard error. Section 4.5 presents a summary of the chapter. This chapter is part of the published paper in Appendix A.4.

With a simplification of notation, I denote by Y_j the *unCL*-summary outcome (labelled $\overline{Y}_j$ in chapter 3) and e_j the *adCL*-summary outcome (previously $\widehat{\overline{e}}_j$).

4.2 Identification assumptions of CL-LATE

To define formally the causal effect of treatment received, I use the *potential outcomes* (POs) framework [44]. I focus on a binary treatment assignment and a level of treatment received at the cluster level, bounded by 0 and 1.

Let $D_{ij} \in \{0, 1\}$ be the treatment received by individual i in cluster j and D_j denote the unadjusted CL treatment received, $D_j = \frac{1}{n_j} \sum_{i=1}^{n_j} D_{ij}$. In the CL adherence settings, D_{ij} is constant within clusters, and therefore D_j is binary. In contrast, when non-adherence is at the individual level, D_j is a continuous measure that varies from 0 to 1, representing the proportion of individuals receiving the active treatment in cluster j, sometimes referred to as treatment intensity [28]. Let also J_0 and J_1 be the number of clusters in the control and active groups, respectively (thus, $J = J_0 + J_1$).

4.2.1 Notation and technical assumptions

I denote by $Y_{ij}(\mathbf{d}_j)$ the PO that would manifest if, possibly contrary to fact, the j-th cluster to which individual unit i belongs receives treatment $\mathbf{d}_j$, a vector of length n_j of 0s and 1s, where we are assuming *no interference between clusters*, i.e. the treatment received by individuals in the j-th cluster are unrelated to the outcome status of individuals in other clusters [13]. *No interference between clusters* is a special case of partial interference, where individual units can be partitioned into groups such that interference does not occur among individuals in different groups but may occur between individuals in the same group [96]. This is commonly assumed in CRTs [13, 30, 96, 97]. A major advantage of CRTs is to reduce interference across clusters but not necessarily across individual units within clusters. The plausibility of this partial *no interference* assumption is enhanced by design (for example, double blinding) or when clusters are well-defined and geographically far apart for instance [8].

Assume *counterfactual consistency*, that is, for $j = 1, \ldots, J$, if $Z_j = z$ then the potential treatment received is $D_{ij} = D_{ij}(z)$ and $Y_{ij} = Y_{ij}(z, D_{ij}(z))$, for all $i = 1, \ldots, n_j$.

4.2.2 Identification assumptions

Assuming *no interference between clusters* and *consistency* allows us to define LATE, the estimand of interest [27].

In the setting considered here where both Z_j and D_{ij} are binary, the vector of *potential treatment received* under alternative random allocation, $\big(D_{ij}(0), D_{ij}(1)\big)$ partitions the individual units in each cluster into four different *compliance* or *adherence classes* [24]: $C_{ij} = n$ (never-takers) if $D_{ij}(0) = D_{ij}(1) = 0$; $C_{ij} = a$ (always-takers) if $D_{ij}(0) = D_{ij}(1) = 1$; $C_{ij} = c$ (compliers) if $D_{ij}(z) = z$ for $z \in \{0, 1\}$; and $C_{ij} = d$ (defiers) if $D_{ij}(z) = 1 - z$ for $z \in \{0, 1\}$. The non-numerical values n, a, c and d, rather than being viewed as values taken by the *adherence classes*, are used to simply help to recognise the *adherence classes*. Any numerical values can be given to n, a, c and d as long as they clearly refer to each *adherence class*. These adherence classes are fixed but unknown for all units.

The estimand of interest here is the so-called *population* LATE, defined as

$$\begin{aligned}\beta &= \mathbb{E}_j\mathbb{E}_i\Big[\Big\{Y_{ij}\big(1, D_{ij}(1)\big) - Y_{ij}\big(0, D_{ij}(0)\big)\Big\}\Big|C_{ij} = c\Big] \\ &= \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\Big\{Y_{ij}\Big(1, D_{ij}(1)\Big) - Y_{ij}\Big(0, D_{ij}(0)\Big)\Big\}\Big\{I\Big(D_{ij}(1) = 1, D_{ij}(0) = 0\Big)\Big\}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} I\Big(D_{ij}(1) = 1, D_{ij}(0) = 0\Big)}\end{aligned} \tag{4.1}$$

This is said to be a *local* causal effect as it is conditional on the stratum of complier individuals. Following [13, 98], we write the cluster version of the corresponding identification assumptions [24] as follows:

(A1) Cluster unconfoundedness : $Z_j \perp\!\!\!\perp D_{ij}(z), Y_{ij}(z, D_{ij}(z)), \quad z \in \{0, 1\}$. This is also known as cluster or group independence or cluster randomisation assumption.
(A2) Exclusion restriction at the individual level : Conditional on the treatment received $D_{ij} = d$, the treatment assignment Z_j had no effect on the outcome. In terms of potential outcomes, we have $Y_{ij}\,(1, d) = Y_{ij}\,(0, d) \quad \forall d \in \{0, 1\}$.
(A3) Instrument relevance: Also referred to as first stage assumption, that is, Z_j is causally associated with treatment received D_{ij}, implying $Z_j \not\perp\!\!\!\perp D_{ij}$.

For point identification of LATE, **(A4) monotonicity** of the treatment mechanism is assumed: $D_{ij}(1) \geq D_{ij}(0)$, often informally referred to as "there are no defiers" [27]. Note that we need to assume this holds at the individual level [98]. For the CL non-adherence setting, where D_{ij} does not vary within clusters, this becomes monotonicity at the cluster level $D_j(1) = 1,\ D_j(0) = 0$.

An extra assumption necessary when using adjusted CL-summary outcomes is that the model used to derive the CL summaries is correctly specified.

In our setting, Z_j is an IV if Z_j fulfils assumptions **(A1)** to **(A3)**. Figure 4.1 summarizes the relationship between the random treatment allocation, the treatment received and outcome variable, assuming that **(A1)** to **(A3)** hold.

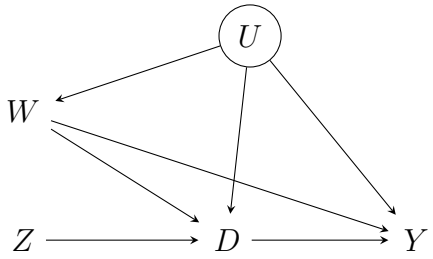


Figure 4.1: Diagram summarising the relationship between random treatment assignment (Z), treatment received (D), measured CL covariate (W), unmeasured covariate (U) and outcome (Y), assuming Z met assumptions **(A1)** to **(A3)**

In randomised trials, assumption **(A1)** holds. **(A3)** and **(A4)** are plausible, particularly in encouragement designs where individuals are encouraged to receive their allocated treatment such as trials where units in the control group are not allowed to have access to the active treatment. **(A2)** may be problematic in trials assessing depression for instance, where patients randomised to the control group may get more depressed for not being offered the active treatment.

4.2.3 Cluster and individual-level non-adherence

The population LATE estimand β can be thought of as a weighted average of the cluster-specific LATE β_j, as follows

$$\beta = \sum_{j=1}^{J} \psi_j \beta_j, \tag{4.2}$$

where

$$\begin{aligned}\beta_j &= \mathbb{E}_i\Big[Y_{ij}\big(1, D_{ij}(1)\big) - Y_{ij}\big(0, D_{ij}(0)\big)\big|C_{ij} = c\Big] \\ &= \frac{1}{n_{c,j}}\sum_{i=1}^{n_j}\Big[\Big\{Y_{ij}\big(1, D_{ij}(1)\big) - Y_{ij}\big(0, D_{ij}(0)\big)\Big\}\Big\{I\big(D_{ij}(1) = 1, D_{ij}(0) = 0\big)\Big\}\Big]\end{aligned} \tag{4.3}$$

where $n_{c,j}$ is the number of individual unit compliers in each cluster j, assumed here to be greater than 0 for all clusters. The weights corresponding to each β_j are $\psi_j = \dfrac{n_{c,j}}{\sum_{j=1}^{J} n_{c,j}}$, that is the number of individual unit compliers in cluster j divided by the total number of individual unit compliers across all clusters. This result is useful when interpreting the estimates obtained using CL summaries. I first note that the cluster-level LATE (CL-LATE) estimator applied to CL summaries, $\hat{\beta}_{\text{CL}}$, does not always correspond to the population LATE β. The former can be expressed as [13]

$$\hat{\beta}_{\text{CL}} = \frac{\mathbb{E}\big[Y_j|Z_j = 1\big] - \mathbb{E}\big[Y_j|Z_j = 0\big]}{\mathbb{E}\big[D_j|Z_j = 1\big] - \mathbb{E}\big[D_j|Z_j = 0\big]} \tag{4.4}$$

In the case where treatment received is at the cluster level (*i.e.* CL adherence), $\hat{\beta}_{\text{CL}}$ can be indeed interpreted as the population LATE. However, when non-adherence varies at the individual level, $\hat{\beta}_{\text{CL}}$ is expressed as [98]

$$\hat{\beta}_{\text{CL}} = \sum_{j=1}^{J} \psi_{\text{CL},j}\hat{\beta}_j \tag{4.5}$$

where the CL-weights are $\psi_{\text{CL},j} = \dfrac{n_{c,j}/n_j}{\sum_j n_{c,j}/n_j}$, *i.e.* the normalised proportion of individual compliers in each cluster [98]. Thus, CL-LATE $\hat{\beta}_{\text{CL}}$ identifies the population LATE β in equation (4.2) only if – (i) the cluster sizes n_j are identical for all clusters, or – (ii) the cluster-specific LATE β_j are homogeneous (*i.e.* constant) across all clusters. If neither of these two conditions is satisfied, then CL-LATE $\hat{\beta}_{\text{CL}}$ cannot be interpreted as an estimator of the population LATE.

When CL covariate adjustment is done, then $\hat{\beta}_{\text{CL}}$ is obtained after standardising (*i.e.* marginalising) the CL-LATE estimator conditional on W_j, over the observed

values of W_j (assuming W_j is categorical and univariate) as follows

$$\hat{\beta}_{\text{CL,adj}} = \frac{\sum_{j=1}^{J} \Big(\mathbb{E}\big[Y_j | Z_j = 1, W_j\big] - \mathbb{E}\big[Y_j | Z_j = 0, W_j\big] \Big) \Pr(W_j)}{\sum_{j=1}^{J} \Big(\mathbb{E}\big[D_j | Z_j = 1, W_j\big] - \mathbb{E}\big[D_j | Z_j = 0, W_j\big] \Big) \Pr(W_j)} \tag{4.6}$$

where $\Pr(W_j)$ is the empirical distribution of the observed Wj.

In the remainder, for IL non-adherence, we assume that the cluster-specific LATE β_j are constant across clusters, while allowing for the cluster sizes n_j to vary. This allows us to interpret $\hat{\beta}_{\text{CL}}$ as an estimator for the population LATE β.

The next section introduces the TSLS method for estimating CL-LATE $\hat{\beta}_{\text{CL}}$.

4.3 TSLS estimation of CL-LATE

TSLS is an IV method that, in the presence of unmeasured confounding, can estimate consistently the causal effect of an exposure under assumptions **(A1)** to **(A3)** plus **(A4)** [24, 27].

TSLS estimation consists of a "first stage" which regresses treatment received on randomised treatment, and a "second stage" which models the outcome on the predicted treatment received [24, 27].

4.3.1 TSLS on CL summaries

I present how to implement TSLS on CL summaries and address issues with heteroscedasticity and inefficiency via weighting and/or HW SEs. I also introduce the small sample degrees of freedom adjustment to accommodate inference based on small samples. In the sections below, I focus on *unCL*-summary outcome as the principles are the same for both *unCL* and *adCL* summaries.

4.3.1.1 Unadjusted CL-LATE

The conditional expectations in equation (4.6) can be estimated via TSLS regression of the CL summaries (referred to as CL-TSLS). CL-TSLS is most easily explained for settings without weights and covariate adjustment. The first stage fits a regression of CL treatment received (or cluster average treatment received) D_j on treatment assigned Z_j. Then, in the second stage, a regression of the CL-summary outcome (either *unCL* or *adCL*) on the predicted treatment received is fitted. Crucially,

both first and second stages must be linear models for the TSLS estimator to be guaranteed consistent [99, 100]. The first and second stage regressions of D_j and Y_j are respectively

$$\begin{aligned} D_j &= \gamma_0 + \gamma_{\mathrm{Z}} Z_j + \xi_{1j} \\ Y_j &= \beta_0 + \beta_{\mathrm{TSLS}} \widehat{D}_j + \xi_{2j} \end{aligned} \tag{4.7}$$

where ξ_{1j} and ξ_{2j} are assumed i.i.d. with mean zero and constant variance, $\xi_{1j} \perp\!\!\!\perp \xi_{2j}$ and $\widehat{D}_j$ are the predicted treatment received from the first stage regression. The estimator $\widehat{\beta}_{\mathrm{TSLS}}$ of CL-LATE β_{CL} is given by [76]

$$\widehat{\beta}_{\mathrm{TSLS}} = \frac{\widehat{\mathrm{Cov}}(Y_j, Z_j)}{\widehat{\mathrm{Cov}}(D_j, Z_j)} = \frac{\widehat{\mathrm{Cov}}(Y_j, Z_j)/\widehat{\mathrm{Var}}(Z_j)}{\widehat{\mathrm{Cov}}(D_j, Z_j)/\widehat{\mathrm{Var}}(Z_j)} \tag{4.8}$$

From equations (4.7) and (4.8), we can see that $\widehat{\beta}_{\mathrm{TSLS}}$ is the ratio of the regression coefficient of Y_j on Z_j to the regression coefficient of D_j on Z_j, γ_{Z}. The estimated asymptotic variance of $\widehat{\beta}_{\mathrm{TSLS}}$ under the assumption of homoscedasticity of residuals in the second stage is [80]

$$\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}) = \widehat{\sigma}^2_{\xi_2} \Big[\sum_{j=1}^{J} D'_j \mathbf{z_j} \Big(\sum_{j=1}^{J} \mathbf{z'_j z_j} \Big)^{-1} \mathbf{z'_j} D_j \Big]^{-1} = \widehat{\sigma}^2_{\xi_2} \Big[\sum_{j=1}^{J} D'_j \widehat{D}_j \Big]^{-1} \tag{4.9}$$

where $\mathbf{z'}_j = \begin{bmatrix} 1 & Z_j \end{bmatrix}$ is the 2×1 vector of regressor and intercept indicator for cluster j, $\widehat{\sigma}^2_{\xi_2}$ is the variance of the residual in the second stage regression. Note that $\begin{bmatrix} \widehat{\gamma}_0 & \widehat{\gamma}_{\mathrm{Z}} \end{bmatrix}' = \Big(\sum_{j=1}^{J} \mathbf{z'}_j \mathbf{z}_j \Big)^{-1} \mathbf{z'}_j D_j$ and $\widehat{D}_j = \mathbf{z}_j \widehat{\gamma}_{\mathrm{Z}}$ from the first stage regression. To account for the prediction error of $\widehat{D}_j$ when estimating $\widehat{\sigma}^2_{\xi_2}$, it is advised not to implement TSLS by hand but rather to use common statistical software packages like Stata and R.

$\widehat{\beta}_{\mathrm{TSLS}}$ is asymptotically normal distributed *i.e.* $\widehat{\beta}_{\mathrm{TSLS}} \sim N(\beta_{\mathrm{TSLS}}, \widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}))$. This asymptotic distribution is often assumed when implementing TSLS in common software like Stata and R.

Heteroscedasticity-robust standard errors

The estimated asymptotic covariance matrix of TSLS estimator $\widehat{\beta}_{IV}$ [80] is $\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}) = J\Big(\sum_{j=1}^{J} D_j'\widehat{D}_j\Big)^{-1}\Big[\sum_{j=1}^{J} D_j'\mathbf{z}_j\Big(\sum_{j=1}^{J}\mathbf{z}'_j\mathbf{z}_j\Big)^{-1}\Big\{\frac{1}{J}\sum_{j=1}^{J}\widehat{\sigma}^2_{\xi_{2j}}\mathbf{z}'_j\mathbf{z}_j\Big\}\Big(\sum_{j=1}^{J}\mathbf{z}'_j\mathbf{z}_j\Big)^{-1}\mathbf{z}'_j D_j\Big]^{-1}\Big(\sum_{j=1}^{J} D_j'\widehat{D}_j\Big)^{-1}$, where $\widehat{\sigma}^2_{\xi_{2j}}$ is the variance of the residual of cluster j in the second step regression. Like in the OLS setting, the HW estimated asymptotic variance of TSLS estimator [70] is

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}) = J\Big(\sum_{j=1}^{J} D_j'\widehat{D}_j\Big)^{-1}\Big[\sum_{j=1}^{J} D_j'\mathbf{z_j}\Big(\sum_{j=1}^{J}\mathbf{z_j'z_j}\Big)^{-1}\Big\{\frac{1}{J}\sum_{j=1}^{J}\widehat{\xi}^2_{2j}\mathbf{z_j'z_j}\Big\}\\ \Big(\sum_{j=1}^{J}\mathbf{z_j'z_j}\Big)^{-1}\mathbf{z_j'}D_j\Big]^{-1}\Big[\sum_{j=1}^{J} D_j'\widehat{D}_j\Big]^{-1}\end{aligned} \tag{4.10}$$

However, HW SEs are recommended but perform poorly in small samples. Testing for conditional heteroscedasticity, using for example the test of Breusch and Pagan [94] or White [70], may guide whether to use weights or not. Unweighted analyses would be preferred if there is no evidence of heteroscedasticity.

Weighted TSLS

When weights are used, each TSLS regression stage is weighted. I refer to this as weighted TSLS (WTSLS). The first and second stage regression equations like in section 3.3.3.3 of chapter 3 are as follows

$$\begin{aligned} D_j\sqrt{\omega_j} &= \gamma_0\sqrt{\omega_j} + \gamma_{\mathrm{Z}}Z_j\sqrt{\omega_j} + \xi_{1j}\sqrt{\omega_j} \\ Y_j\sqrt{\omega_j} &= \beta_0\sqrt{\omega_j} + \beta_{\mathrm{TSLS}}\widehat{D}_j\sqrt{\omega_j} + \xi_{2j}\sqrt{\omega_j} \end{aligned} \tag{4.11}$$

where $\widetilde{\xi}_{1j} = \xi_{1j}\sqrt{\omega_j}$ and $\widetilde{\xi}_{2j} = \xi_{2j}\sqrt{\omega_j}$ are the error terms with mean zero and constant variance, and $\widetilde{\xi}_{1j} \perp\!\!\!\perp \widetilde{\xi}_{2j}$.

Let denote $Y_j\sqrt{\omega_j}$ by $\widetilde{Y}_j$, $Z_j\sqrt{\omega_j}$ by $\widetilde{Z}_j$, $D_j\sqrt{\omega_j}$ by $\widetilde{D}_j$, $\widehat{D}_j\sqrt{\omega_j}$ by $\widetilde{\widehat{D}}_j$ and Ω_j a variable whose values are $\sqrt{\omega_j}$. The 2×1 vector $\mathbf{z}'_j = \begin{bmatrix}1 & Z_j\end{bmatrix}$ for cluster j becomes $\widetilde{\mathbf{z}}_{\mathbf{j}} = \begin{bmatrix}\Omega_j & \widetilde{Z}_j\end{bmatrix}$. The unadjusted WTSLS CL-LATE estimator is

$$\widehat{\beta}_{\mathrm{TSLS}} = \frac{\widehat{\mathrm{Cov}}(\widetilde{Y}_j, \widetilde{Z}_j)}{\widehat{\mathrm{Cov}}(\widetilde{D}_j, \widetilde{Z}_j)} = \frac{\widehat{\mathrm{Cov}}(\widetilde{Y}_j, \widetilde{Z}_j)/\widehat{\mathrm{Var}}(\widetilde{Z}_j)}{\widehat{\mathrm{Cov}}(\widetilde{D}_j, \widetilde{Z}_j)/\widehat{\mathrm{Var}}(\widetilde{Z}_j)} \tag{4.12}$$

Similar to (4.9), the estimated WTSLS asymptotic variance of $\widehat{\beta}_{\text{TSLS}}$ is

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{TSLS}}) = \widehat{\sigma}^2_{\widetilde{\xi}_2}\Big[\sum_{j=1}^{J}\widetilde{D}_j'\widetilde{\mathbf{z}}_j\Big(\sum_{j=1}^{J}\widetilde{\mathbf{z}}_j'\widetilde{\mathbf{z}}_j\Big)^{-1}\widetilde{\mathbf{z}}_j'\widetilde{D}_j\Big]^{-1} = \widehat{\sigma}^2_{\widetilde{\xi}_2}\Big[\sum_{j=1}^{J}\widetilde{D}_j'\widetilde{\widehat{D}}_j\Big]^{-1} \tag{4.13}$$

Small sample degrees of freedom

The asymptotic normal distribution used by default in standard statistical software to test and construct the 95% confidence interval of $\widehat{\beta}_{\text{TSLS}}$ may not be adequate, especially with small number of clusters. Rather than relying on asymptotic properties, the exact $t-$distribution of $\widehat{\beta}_{\text{TSLS}}$ can be used instead. This is so-called "small sample degrees of freedom" (SSDF) adjustment. When SSDF correction is done, $\widehat{\beta}_{\text{TSLS}} \sim t(J-2)$, where the general form of its estimated variance is

$$\begin{aligned}\widehat{\text{Var}}(\beta_{\text{TSLS}}) = J\Big(\sum_{j=1}^{J}D_j'\widehat{D}_j\Big)^{-1}\Big[\sum_{j=1}^{J}D_j'\mathbf{z_j}\Big(\sum_{j=1}^{J}\mathbf{z_j'z_j}\Big)^{-1}\Big\{\frac{1}{J-2}\sum_{j=1}^{J}\widehat{\sigma}^2_{\xi_{2j}}\mathbf{z_j'z_j}\Big\} \\ \Big(\sum_{j=1}^{J}\mathbf{z_j'z_j}\Big)^{-1}\mathbf{z_j'}D_j\Big]^{-1}\Big[\sum_{j=1}^{J}D_j'\widehat{D}_j\Big]^{-1}\end{aligned} \tag{4.14}$$

The SSDF correction consists of dividing the residual variance by the sample size minus the number of parameters in the second stage regression, here $J-2$.

4.3.1.2 Covariate-adjusted CL-LATE

Trial investigators often plan to adjust for baseline covariates when estimating the ITT effect and also the causal effect, if of interest. It is possible to obtain covariate-adjusted CL-LATE estimate from TSLS by including the covariates in both first and second stage regressions as follows

$$\begin{aligned} D_j &= \gamma_0 + \gamma_{\text{Z}}Z_j + \gamma_{\text{W}}W_j + \xi_{1j} \\ Y_j &= \beta_0 + \beta_{\text{IV,adj}}\widehat{D}_j + \beta_{\text{W}}W_j + \xi_{2j} \end{aligned} \tag{4.15}$$

where ξ_{1j} and ξ_{2j} are assumed i.i.d. with mean zero and constant variance, ξ_{1j} and ξ_{2j} are uncorrelated, and $\widehat{D}_j$ are the predicted treatment received from first stage regression. Analogously to the "regression anatomy" [76, 83], introduced in section

3.7, the estimator $\widehat{\beta}_{\text{IV,adj}}$ of CL-LATE β_{CL} can be expressed as [76]

$$\widehat{\beta}_{\text{IV,adj}} = \frac{\widehat{\text{Cov}}(Y_j, \epsilon_{z_j})}{\widehat{\text{Cov}}(D_j, \epsilon_{z_j})} = \frac{\widehat{\text{Cov}}(Y_j, \epsilon_{z_j})/\widehat{\text{Var}}(\epsilon_{z_j})}{\widehat{\text{Cov}}(D_j, \epsilon_{z_j})/\widehat{\text{Var}}(\epsilon_{z_j})} \tag{4.16}$$

where ϵ_{z_j} are the residuals from regressing Z_j on W_j. $\widehat{\beta}_{\text{IV,adj}}$ is the ratio of the regression coefficient of Y_j on ϵ_{z_j} to the regression coefficient of D_j on ϵ_{z_j}. Note that equation (4.16) holds when there is a single IV (here, Z_j) and a single exposure of interest (here, D_j) like in randomised trials setting [76].

ϵ_{z_j} can be viewed as an IV improved by eliminating any possible association between Z_j and measured covariates W_j. However, as mentioned in section 4.2.3, adjusting for covariates in TSLS when the IV is a randomised treatment does not change the specification of the LATE estimator. This is obvious in equation (4.16) as $\epsilon_{z_j} = Z_j - \mathbb{E}(Z_j)$ because $Z_j \perp\!\!\!\perp W_j$ (*i.e.* Z_j independent of W_j) by randomisation. Consider the linear regression model $Z_j = \delta_0 + \delta_{\text{W}} W_j + \epsilon_{z_j}$. $Z_j \perp\!\!\!\perp W_j$ implies that $\delta_{\text{W}} = 0$ and the predicted value of Z_j is $\widehat{Z}_j = \mathbb{E}(Z_j)$. Thus, $\epsilon_{z_j} = Z_j - \mathbb{E}(Z_j)$. It follows that $\text{Cov}(Y_j, \epsilon_{z_j}) = \text{Cov}(Y_j, Z_j - \mathbb{E}(Z_j)) = \text{Cov}(Y_j, Z_j)$ and likewise $\text{Cov}(D_j, \epsilon_{z_j}) = \text{Cov}(D_j, Z_j - \mathbb{E}(Z_j)) = \text{Cov}(D_j, Z_j)$. Hence, $\widehat{\beta}_{\text{IV,adj}} = \frac{\widehat{\text{Cov}}(Y_j, Z_j)}{\widehat{\text{Cov}}(D_j, Z_j)} = \widehat{\beta}_{\text{TSLS}}$. This supports the reason why it is not necessary to adjust for covariates in TSLS when there is a valid IV as we do not expect to gain any bias reduction [101]. The equivalence of $\widehat{\beta}_{\text{IV,adj}}$ and $\widehat{\beta}_{\text{TSLS}}$ may not be apparent in finite randomised trial samples because of possible baseline covariates imbalance.

However, there will be gain in precision if W_j is associated with Y_j and D_j. Though not obvious in equation (4.16), adjusting for W_j when associated with D_j and Y_j reduces both first and second stage residual variances.

The estimated asymptotic variance of $\widehat{\beta}_{\text{TSLS}}$ assuming homoscedasticity [80] is

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{TSLS}}) = \widehat{\sigma}^2_{\xi_2} \Big[\sum_{j=1}^{J} D'_j \mathbf{z}_j \Big(\sum_{j=1}^{J} \mathbf{z}'_j \mathbf{z}_j \Big)^{-1} \mathbf{z}'_j D_j \Big]^{-1} = \widehat{\sigma}^2_{\xi_{2j}} \Big[\sum_{j=1}^{J} D'_j \widehat{D}_j \Big]^{-1} \tag{4.17}$$

where $\mathbf{z}'_j = \begin{bmatrix} 1 & Z_j & W_j \end{bmatrix}$ is the $p \times 1$ vector of regressors and intercept indicator (where W_j consists of $(p-2)$ covariates) for cluster j, $\widehat{\sigma}^2_{\xi_2}$ is the residual variance in

the second stage regression.

Heteroscedasticity-robust standard errors

The HW estimated asymptotic variance of $\widehat{\text{Var}}(\beta_{\text{IV,adj}})$ is of similar form to (4.10) [70], except that the 2×1 vector of regressors and intercept indicator, $\mathbf{z'}_j = \begin{bmatrix} 1 & Z_j \end{bmatrix}$, is replaced with the $p \times 1$ vector of regressors and intercept indicator for cluster j $\mathbf{z'}_j = \begin{bmatrix} 1 & Z_j & W_j \end{bmatrix}$.

Weighted TSLS

WTSLS with covariate adjustment is similar to the unadjusted WTSLS, except that the covariates W_j are included in each regression stage and $\mathbf{z'}_j = \begin{bmatrix} \Omega_j & \widetilde{Z}_j \end{bmatrix}$ becomes the $p \times 1$ vector $\mathbf{z'}_j = \begin{bmatrix} \Omega_j & \widetilde{Z}_j & \widetilde{W}_j \end{bmatrix}$, where $\widetilde{W}_j = W_j \sqrt{\omega_j}$ and W_j represents $(p-2)$ covariates. The covariate-adjusted WTSLS CL-LATE estimator is

$$\widehat{\beta}_{\text{IV,adj}} = \frac{\widehat{\text{Cov}}(Y_j, \epsilon_{\widetilde{z}_j})}{\widehat{\text{Cov}}(D_j, \epsilon_{\widetilde{z}_j})} \tag{4.18}$$

where $\epsilon_{\widetilde{z}_j}$ are the residuals from regressing $\widetilde{Z}_j$ on $\widetilde{W}_j$ without intercept. Similarly to (4.9), the estimated WTSLS asymptotic variance of $\widehat{\beta}_{\text{IV,adj}}$ has a similar form as 4.13, where $\mathbf{z'}_j = \begin{bmatrix} 1 & Z_j & W_j \end{bmatrix}$ is the $p \times 1$ vector of regressors and intercept indicator for cluster j.

Small sample degrees of freedom

SSDF correction procedure is the same as in 4.3.1.1. When SSDF is done in the presence of $(p-2)$ covariates (Z_j not included), $\widehat{\beta}_{\text{IV,adj}} \sim t(J-p)$.

4.4 Schochet-Chiang approach

The Schochet-Chiang's estimation of CL-LATE is based on the Wald estimator [31], which I denote here by $\widehat{\beta}_{\text{CL,Wald}}$. I introduce the Wald estimator and its standard error as suggested by Schochet and Chiang [13].

4.4.1 Wald estimator

The Wald estimator [31] is a ratio of the ITT effect on the outcome (effect of Z on Y) to the ITT effect on treatment (effect of Z on D). The basis of the Wald estimator [31] is equation (4.4) where Z_j is binary and there is no covariate adjustment. $\widehat{\beta}_{\text{CL,Wald}}$ is a simple and consistent estimator of β_{CL} [102]. Schochet and Chiang [13] used OLS

regressions to estimate the ITT effects on the *unCL* summaries Y_j and on the *unCL* summaries D_j when adherence is at the individual level or simply the treatment received for CL adherence. Only CL covariate adjustment is valid [13, 76].

When there is no covariate adjustment, the following equations are used to obtain the Wald estimator:

$$Y_j = \beta_0 + \beta_{\mathrm{Z}} Z_j + \eta_{1_j} \tag{4.19}$$

$$D_j = \gamma_0 + \gamma_{\mathrm{Z}} Z_j + \eta_{2_j} \tag{4.20}$$

Covariate adjustment is done as shown in equations (4.21) and (4.22) as follows

$$Y_j = \beta_0 + \beta_{\mathrm{Z}} Z_j + \beta_{\mathrm{W}} W_j + \eta_{1_j} \tag{4.21}$$

$$D_j = \gamma_0 + \gamma_{\mathrm{Z}} Z_j + \gamma_{\mathrm{W}} W_j + \eta_{2_j} \tag{4.22}$$

where η_{1_j} and η_{2_j} are *i.i.d* normal residuals with mean 0 and are independent of each other. The Wald estimator of β is

$$\widehat{\beta}_{\mathrm{CL,Wald}} = \frac{\widehat{\beta}_{\mathrm{Z}}}{\widehat{\gamma}_{\mathrm{Z}}}. \tag{4.23}$$

$\widehat{\beta}_{\mathrm{CL,Wald}}$ is a non-linear combination of two estimators and therefore obtaining an analytic form of its variance may entail some approximations. I present two ways of estimating the standard error of $\widehat{\beta}_{\mathrm{CL,Wald}}$. These include the traditional and the Schochet-Chiang approaches.

A drawback of estimating the standard error of $\widehat{\beta}_{\mathrm{CL,Wald}}$ using the *unCL* summaries Y_j and D_j is the inability to appropriately account for the influence of X_{ij} on the precision of $\widehat{\beta}_{\mathrm{CL,Wald}}$.

4.4.2 Traditional standard errors

An analytic form of the variance of a non-linear combination of estimators is commonly obtained using the first order approximation of Taylor series expansion, commonly known as the "Delta" method [103]. The first order approximation of Taylor series expansion for $\mathrm{Var}(\widehat{\beta}_{\mathrm{CL,Wald}})$ is

$$\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\beta}_{\mathrm{CL,Wald}}) = \frac{\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\beta}_{\mathrm{Z}})}{\widehat{\gamma}^2_{\mathrm{Z}}} + \widehat{\beta}^2_{\mathrm{CL,Wald}} \frac{\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\gamma}_{\mathrm{Z}})}{\widehat{\gamma}^2_{\mathrm{Z}}} - 2\widehat{\beta}_{\mathrm{CL,Wald}} \frac{\widehat{\mathrm{C}}\mathrm{ov}(\widehat{\beta}_{\hat{\mathrm{Z}}}, \hat{\gamma}_{\mathrm{Z}})}{\widehat{\gamma}^2_{\mathrm{Z}}} \tag{4.24}$$

The variance of $\widehat{\beta}_{\text{CL,Wald}}$ is traditionally estimated assuming that the denominator $\widehat{\gamma}_{\text{Z}}$ is estimated without error [13]. Thus, $\widehat{\text{Var}}(\widehat{\gamma}_{\text{Z}}) = 0$ and $\widehat{\text{Cov}}(\widehat{\beta}_{\text{Z}}, \hat{\gamma}_{\text{Z}}) = 0$ and equation (4.24) simplifies to [13, 98]

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{CL,Wald}}) = \frac{\widehat{\text{Var}}(\widehat{\beta}_{\text{Z}})}{\widehat{\gamma}_{\text{Z}}^2} = \frac{J\widehat{\sigma}_{\eta_1}^2}{J_0 J_1 \widehat{\gamma}_{\text{Z}}^2} \tag{4.25}$$

where $\widehat{\sigma}_{\eta_1}^2 = \frac{1}{J-2}\sum_{i=1}^{J}(Y_j - \widehat{Y}_j)^2$ is the residual variance from equation (4.19), $\widehat{Y}_j$ is the predicted value of *unCL*-summary outcome for cluster j and $J-2$ are the degrees of freedom (here, number of clusters minus number of parameters in equation (4.19)). Here, it is assumed that the residual variance is the same variance in the control and active groups. This variance homogeneity assumption can be easily relaxed. In the presence of CL covariates adjustment like in equation (4.21), the degrees of freedom are equal to $J - p$ where p is the number of CL regressors (Z_j included) plus the intercept. The standard error of $\widehat{\beta}_{\text{CL,Wald}}$ is then obtained by taking the square root of equation (4.25).

The "Delta" method has been shown to perform poorly, particularly when proportion of adherent units is low [98, 104].

4.4.3 Schochet-Chiang standard errors

Schochet and Chiang [13] estimated the variance of $\widehat{\beta}_{\text{CL,Wald}}$ using equation (4.24). This allows us to account for the estimation error in both numerator ($\widehat{\beta}_{\text{Z}}$) and denominator ($\widehat{\gamma}_{\text{Z}}$) of the Wald estimator. The second term of equation (4.24) accounts for the uncertainty of $\widehat{\gamma}_{\text{Z}}$ and the third term involves the covariance of $\widehat{\beta}_{\text{Z}}$ and $\widehat{\gamma}_{\text{Z}}$. It is possible to account for potential heteroscedasticity likely to occur from varying precision of CL summaries, while estimating the variance of the Wald estimator as in [13]. However, Schochet and Chiang found very similar results whether allowing for weights or not when estimating the corrected variance [13].

Like in [13], I assume unequal variances of CL-summary treatment received D_j across trial groups as this seems plausible in randomised experiments; for example, in one-sided non-adherence where none of the units in the control group receives the

active treatment. The variance of $\widehat{\gamma}_{\mathrm{z}}$, without weighting, is given by [13]

$$\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\gamma}_{\mathrm{z}}) = \frac{\widehat{S}_0^2}{J_0(J_0-p)} + \frac{\widehat{S}_1^2}{J_1(J_1-p)} \tag{4.26}$$

where $\widehat{S}_0^2 = \frac{1}{J_0-p}\sum_{i=1}^{J_0}(1-Z_j)(D_j-\widehat{D}_j)^2$ and $\widehat{S}_1^2 = \frac{1}{J_1-p}\sum_{i=1}^{J_1}Z_j(D_j-\widehat{D}_j)^2$, and $\widehat{D}_j$ is the predicted value of D_j from equation (4.20) if no covariate adjustment and (4.22) otherwise. A consistent estimator of the covariance of $\widehat{\beta}_{\mathrm{z}}$ and $\widehat{\gamma}_{\mathrm{z}}$ is [13]

$$\widehat{\mathrm{C}}\mathrm{ov}(\widehat{\beta}_{\mathrm{z}},\widehat{\gamma}_{\mathrm{z}}) = \frac{\sum_{i=1}^{J_0}(1-Z_j)(Y_j-\widehat{Y}_j)(D_j-\widehat{D}_j)}{J_0(J_0-p)} + \frac{\sum_{i=1}^{J_1}Z_j(Y_j-\widehat{Y}_j)(D_j-\widehat{D}_j)}{J_1(J_1-p)} \tag{4.27}$$

Thus, after inserting equations (4.25), (4.26) and (4.27) in equation (4.24), the estimated asymptotic variance of $\widehat{\beta}_{\mathrm{CL,Wald}}$ is as follows

$$\begin{aligned}\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\beta}_{\mathrm{CL,Wald}}) = \frac{J\widehat{\sigma}^2_{\eta_1}}{J_0J_1\widehat{\gamma}_{\mathrm{z}}^2} + \frac{\widehat{\beta}^2_{\mathrm{CL,Wald}}}{\widehat{\gamma}_{\mathrm{z}}^2}\Big(\frac{\sum_{i=1}^{J_0}(1-Z_j)(D_j-\widehat{D}_j)^2}{J_0(J_0-p)} + \frac{\sum_{i=1}^{J_1}Z_j(D_j-\widehat{D}_j)^2}{J_1(J_1-p)}\Big) - \\ 2\frac{\widehat{\beta}_{\mathrm{CL,Wald}}}{\widehat{\gamma}_{\mathrm{z}}^2}\Big(\frac{\sum_{i=1}^{J_0}(1-Z_j)(Y_j-\widehat{Y}_j)(D_j-\widehat{D}_j)}{J_0(J_0-p)} + \frac{\sum_{i=1}^{J_1}Z_j(Y_j-\widehat{Y}_j)(D_j-\widehat{D}_j)}{J_1(J_1-p)}\Big)\end{aligned} \tag{4.28}$$

In large samples, $\frac{\widehat{\beta}_{\mathrm{CL,Wald}}}{\widehat{\mathrm{V}}\mathrm{ar}(\widehat{\beta}_{\mathrm{CL,Wald}})}$ follows a standard normal distribution [98].

The Wald and TSLS estimators are equivalent in the absence of covariate adjustment [80]. When applied to CL summaries without weighting, the point estimates from the Schochet-Chiang method are equivalent to those from TSLS when there is no covariate adjustment.

4.5 Summary

I presented CL-TSLS and the Wald estimator with Schochet-Chiang SEs for estimating CL-LATE. In the presence of non-adherence, randomised treatment can be used as an instrument to perform TSLS estimation of CL-LATE, that is the causal effect of actually receiving the treatment on the CL-mean outcome. Covariates adjustment is not required because of randomisation being a valid instrument, but rather aims to improve precision.

CL-summary analyses may lead to invalid inferences because of likely heteroscedasticity due to the varying clusters size. Estimation by WLS, where the weights are

defined either by the CS or by the MV weights can improve efficiency. The use of weights and/or HW SEs help tackling heteroscedasticity, if present. However, HW SEs perform poorly in small samples [71]. Testing for conditional heteroscedasticity, using for example the test of Breusch and Pagan [94] or White [70], may guide whether to use weights or not. Unweighted analyses would be preferred if there is no evidence of heteroscedasticity. Inference in TSLS estimation often relies on normality from large sample properties. Therefore, it may be necessary to use the exact $t-$distribution in small samples by choosing the small sample degree of freedom adjustment, available in standard software.

The Wald estimator for CL-LATE is a ratio where the numerator and denominator are obtained via OLS estimations. The standard error of the ratio is estimated using the first order approximation of Taylor series expansion as per Schochet-Chiang [13]. The point estimates from the Wald estimator are equivalent to those from TSLS in the absence of covariate adjustment.

Chapter 5.

Simulation study of cluster-level LATE estimation in CRTs

5.1 Introduction

This chapter investigates, via simulations, the finite sample performance of TSLS and the Wald estimator with Schochet-Chiang SEs for CL-LATE introduced in chapter 4. I evaluate their empirical bias and coverage. The simulation study considers one-sided non-adherence CRTs of different sizes and where non-adherence is either at the cluster or individual level. I allow for various effect sizes of cluster-level and individual-level variables on the outcome and the treatment received.

The chapter is organized as follows. Section 5.2 introduces the data generating process. Section 5.3 summarises the analysis and criteria used to assess the methods performance. Section 5.4 presents the performance of CL-TSLS and the Wald estimator with Schochet-Chiang SEs. Some additional investigations are done and presented in section 5.5. Finally, section 5.6 summarizes the chapter.

5.2 Data generating process

I simulate CRT with individual-level data where there is one-sided non-adherence at either the cluster or individual level. With a fixed expected total sample size $n = 1000$ individuals, I vary the number of clusters J and the average cluster size n_j. The marginal ICC of Y also takes two values. The effect of cluster-level and individual-level variables on the outcome and the treatment received also varies, so that the strength of the confounding (observed and unobserved) is either low or high, while the value of the true LATE also has two levels *i.e.* low (0.1 standard deviation of the outcome) or modest (0.4 standard deviation of the outcome). The cluster-

level and individual-level covariates are confounders of the relationship between the treatment received and the outcome. The unobserved confounding is introduced by omitting either or both covariates from the analysis. In other words, there is no unobserved confounding when both cluster-level and individual-level covariates are included in the analysis. Table 5.1 summarises the factorial design and the values taken by the different levels.

More specifically, I simulate cluster randomised treatment $Z_j \sim \text{Bern}(0.5)$ and two independent baseline covariates, a cluster-level covariate $W_j \sim N(0, \sigma^2_{\text{w}})$ and individual-level covariate $X_{ij} \sim N(0, \sigma^2_{\text{x}})$ with a moderate ICC $\rho_{\text{x}} = 0.05$, and $\sigma^2_{\text{w}} = \sigma^2_{\text{x}} = 0.08$. I chose $\rho_{\text{x}} = 0.05$ as Kul *et al.* [105] reported a median ICC value (1st-3rd quartiles) of 0.043 (0.026-0.052) for baseline characteristics in CRTs in health care.

I generate a binary adherence class indicator variable C_{ij}, which is considered as latent. I present below how this indicator variable is generated for cluster-level and individual-level adherence settings.

For settings where adherence is at the cluster level, the adherence class indicator variable is constant within clusters, under the following model

$$C_{ij} = C_j \sim \text{Bern}(\pi_j) \quad \text{with} \quad \pi_j = P(C_j = 1)$$
$$\text{logit}(\pi_j) = \lambda_0 + \lambda_{\text{w}} W_j,$$

with $\lambda_{\text{w}} = 0.05$ equivalent to an odds ratio OR ≈ 1.05 per unit increase in W (denoted "small effect") and $\lambda_{\text{w}} = 0.7$ equivalent to OR ≈ 2 ("large effect"). I chose $\pi_j = 0.6$ on average, as the systematic review conducted in chapter 2 shows the median proportion of non-adherent clusters in the active group to be 44.8%.

Table 5.1: Factorial design of the data generating processes and values taken by the parameters in the simulations

Parameter	Label	Level	Value
CRT size			
n	Total number of individuals	Moderate	≈ 1000
J	Number of clusters and	Moderate clusters	$J = 50, n_j \sim \text{Poi}(20)$
n_j	individuals per cluster	Few large clusters	$J = 10, n_j \sim \text{Poi}(100)$
Baseline variables			
W_j	Cluster-level variable	-	$W_j \sim N(0, 0.08)$
ρ_{X}	ICC for X_{ij}	Moderate	0.05
X_{ij}	Individual-level variable	-	$X_{ij} = X_j + e_{ij}$, $X_j \sim N(0, 0.004)$, $e_{ij} \sim N(0, 0.076)$
Adherence to treatment			
π	Expected probability of adherence	Moderate	0.60 (cluster-level), 0.85 (individual-level)
λ_{W}, λ_{X}	W_j and X_{ij} effects on log odds of adherence	Small, Large	$\lambda_{\text{W}} = \lambda_{\text{X}} = 0.05$, $\lambda_{\text{W}} = \lambda_{\text{X}} = 0.70$
C_j	Cluster-level adherence class	-	$\text{Bern}[\texttt{expit}(\lambda_0 + \lambda_{\text{W}} W_j)]$
C_{ij}	Individual-level adherence class	-	$\text{Bern}[\texttt{expit}(\lambda_0 + \lambda_{\text{W}} W_j + \lambda_{\text{X}} X_{ij} + \zeta_j)]$
ζ_j	Cluster-level random effects	-	$\zeta_j \sim N(0, \pi^2/3)$
ρ_{C}	ICC for C_{ij}	Moderate	0.50
Outcome			
β_0, β_{C}			β_0=0, β_{C}=0
β_{W}, β_{X}	W_j and X_{ij} effects on outcome Y_{ij}	Small, Modest	$\beta_{\text{W}} = \beta_{\text{X}}$=0.1 SD, $\beta_{\text{W}} = \beta_{\text{X}}$=0.4 SD
β_{CZ}	True LATE	Small, Modest	0.1 SD, 0.4 SD
ρ_{Y}	ICC for Y_{ij}	Small, Large	0.05, 0.20

[a] SD: standard deviation of the outcome Y, $\sigma = 1$.

For individual-level adherence, the adherence class indicator variable is generated using the model as follows

$$C_{ij} \sim \text{Bern}(\pi_{ij}) \quad \text{with} \quad \pi_{ij} = P(C_{ij} = 1)$$
$$\text{logit}(\pi_{ij}) = \lambda_0 + \lambda_{\text{W}} W_j + \lambda_{\text{X}} X_{ij} + \zeta_j$$
$$\zeta_j \sim N\left(0, \sigma_\zeta^2\right)$$

with $\sigma_\zeta^2 = \pi^2/3$, so that the ICC for compliance is $\rho_{\text{C}} = \sigma_\zeta^2/(\sigma_\zeta^2 + \pi^2/3) = 0.50$. The treatment received at the individual level is derived as $D_{ij} = Z_j C_{ij}$, so that as applied to one-sided non-adherence, those individuals in clusters randomly allocated to control always have control treatment, but those in clusters randomised to the active intervention can switch to the control treatment, depending on their adherence class. I chose an expected value of π_{ij} of 0.8, as the systematic review conducted in chapter 2 shows that the median proportion of non-adherent individuals at the cluster level is 15% in the active group.

I finally generate continuous outcome Y_{ij}, under the ER assumption,

$$Y_{ij} = \beta_0 + \beta_{\text{C}} C_{ij} + \beta_{\text{CZ}} C_{ij} Z_j + \beta_{\text{W}} W_j + \beta_{\text{X}} X_{ij} + \upsilon_j + \epsilon_{ij} \tag{5.1}$$

with $\upsilon_j \sim N(0, \sigma_\upsilon^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where the values for σ_υ^2 and σ_ϵ^2 are chosen such that the marginal ICC for Y has the corresponding value according to the simulated scenario, given that $\text{Var}(Y_{ij}) = \sigma^2 = 1$. The choice of the parameters' values is reported in Table 5.1. Note that the treatment effects are assumed to be homogeneous within principal strata but heterogeneous across principal strata *i.e.* the treatment effects for adherent units are different from those of non-adherent units but are similar across units within each of those adherence classes.

Inclusion criteria for simulated datasets

I need the data generating process to result in randomised treatment Z that is a valid IV *i.e.* fulfilling assumptions **(A1)** to **(A3)** introduced in section 4.2.2. However, with the choices made, some simulated CRT datasets may result in weak instruments, especially for cluster-level non-adherence settings, with only 5 clusters per trial arm and an expected proportion of non-adherent clusters set at 40%.

Thus, after creating each dataset, I perform an unadjusted first stage regression of D_j on Z_j, and reject simulated datasets where the resulting $F-$statistic is < 10 as per Staiger & Stock's rule of thumb for weak instruments [106]. I continue this process until I get 2500 datasets with valid IV per scenario. For settings where non-adherence is at the cluster level and only 5 clusters are assigned to each trial group, about half of the generated CRTs result in weak instruments. This combination of settings was included as an extreme scenario, and the rejection of simulated CRTs where random treatment assignment is weakly or not associated with the treatment received, was to ensure that we only simulated well conducted trials where a well-argued case for a LATE analysis can be made. Random allocation not being associated with treatment received is indicative of a failure of the study conduct and/or the intervention being considered unacceptable by the participants. In such extreme cases, a LATE analysis is not recommended. In practice, trials are expected to be well conducted and therefore the simulation study is relevant for most CRT settings.

The R code used to generate the datasets and the Stata code for CL summary-based data analysis are shown in appendices A.8 and A.9, respectively.

5.3 Analysis and performance criteria

I consider the TSLS estimation using CL summaries and the Wald estimator with Schochet-Chiang SEs introduced in chapter 4. Analyses are performed using Stata 15. Details of the CL summary-based analysis code can be found in appendix A.9. A summary of the analysis scenario is given in Table 5.2.

5.3.1 TSLS and Wald estimator with Schochet-Chiang SEs

Recall that for TSLS, estimation in each scenario involves using unadjusted CL summary of treatment received in the first-stage, and either unadjusted or individual-level variable adjusted CL summary outcomes, for the second stage. Each regression in the TSLS was fitted via OLS or WLS, the latter with either CS or MV weights. I also consider TSLS where each stage model is either unadjusted or adjusted for a cluster-level variable. Finally, I obtain SEs assuming homoscedasticity or het-

eroscedasticity, and SSDF-based or normal approximation-based CIs. As for the Wald estimator with Schochet-Chiang SEs, only the unadjusted CL summaries of the outcome and treatment received are analysed via OLS.

Table 5.2: Overview of TSLS and Wald estimator with Schochet-Chiang SEs of CL-LATE and inference strategies used in the simulation study

Methods	**Analyses features**			
TSLS	CL outcome	Unadjusted	Adjusted for X_{ij}	
	Adjusted for W_j	No	Yes	
	Weights	None (i.e. OLS)	*CS*	*MV*
	SE estimation	Normal theory	HW	
	SSDF correction	No	Yes	
Schochet-Chiang	CL outcome and CL treatment received	Unadjusted		
	Adjusted for W_j	No	Yes	
	Adjusted for X_{ij}	No	No	
	Weights	None (i.e. OLS)		
	SE estimation	First order approximation of Taylor series expansion		

CL: cluster level; HW: Huber-White; CS weights: cluster-size weights; MV weights: minimum variance weights; SE: standard error; SSDF: Small sample degrees of freedom correction.

5.3.2 Performance criteria

The performance criteria used are empirical bias and coverage rates of the 95% CIs over the 2,500 replicate datasets per scenario. For the bias, I construct a 95% CI using its Monte Carlo Error (MCE). The coverage rate sampling error given the size of the simulation results in a valid range between 94.1% and 95.9%.

Let the mean of the estimated LATE across the replicate datasets in each scenario, indexed by $l = 1, ..., L$, with $L = 2\ 500$ be $\bar{\hat{\beta}}_{IV} = \frac{1}{L}\sum_{l=1}^{L}\hat{\beta}_{IV_l}$. The following criteria were used to assess the performance of the methods investigated

(a) **Empirical bias:** estimated by $\bar{\hat{\beta}}_{IV} - \beta_{\text{CZ}}$.

(b) **MCE of empirical bias** $= \sqrt{\sum_{l=1}^{L}\left(\widehat{\beta}_{IV_l} - \bar{\hat{\beta}}_{IV}\right)^2/[L(L-1)]}$.

(c) **Coverage rate of the nominal of 95% CIs** $\frac{1}{L}\sum_{l=1}^{L} I\left(|\hat{\beta}_{IV_l} - \beta_{\text{CZ}}| < 1.96 s_i\right)$, where s_i denotes the model-based SE for $\hat{\beta}_{IV_l}$. The MCE of coverage is given by $\sqrt{\sum_{l=1}^{L}(0.95)(0.05)/L}$, which means that the expected range of the nominal 95% CIs is between 94.1% and 95.9%.

5.4 Results

I present the results for TSLS and the Wald estimator with Schochet-Chiang SEs by plotting the empirical bias with the MCE-based CIs. The valid range for the coverage rate at the 95% nominal confidence interval is represented by horizontal dashed lines.

5.4.1 TSLS estimation

Figure 5.1 and Figure 5.2 report the empirical bias and 95% CI coverage resulting from each of the different CL-TSLS estimators, when adherence is at cluster or individual level respectively, and for scenarios where the true LATE is modest (0.4 SD). The corresponding figures for small true LATE are in Figures 5.3 and 5.4. The scenarios considered here only include clusters with similar size.

Each figure reports results where $J = 10$ (Panel A, top) or $J = 50$ (Panel B), and with the ICC for Y, ρ_{Y}, is either small (first three columns) or large (last three columns). In each cell, the results for alternative combinations of TSLS (unadjusted/adjusted for W_j) applied to unCL or adCL outcomes are plotted along the horizontal axis. The different data generation scenarios are identified by $*, +, \times$, and $\circ$, corresponding to varying strengths of the effects of X and W (considered as either observed or unobserved confounding) on Y.

The CL–TSLS estimators show some finite sample bias in settings where the number of clusters is small (J=10, Panel A), regardless of whether the non-adherence was at the cluster or individual level and whether the CL summary for Y was adjusted or unadjusted or W_j was included or not in the TSLS regressions. However, the Monte-Carlo error CIs includes 0 in many settings. The bias is more severe when the ICC for Y is larger (right hand side of each Figure), especially if the number of clusters is small (Panel A). The bias is somewhat attenuated when we adjust for W_j in the TSLS, and the non-adherence is at the cluster-level (Figures 5.1 and 5.3). In contrast, for settings with individual-level non-adherence, this adjustment instead increases the bias, especially if W has only a small confounding effect. In these scenarios, the estimates exhibit a small but statistically significant bias, which

disappears when the number of clusters is larger (Figures 5.2 and 5.4). In general, the bias is not affected by the choice of weighting strategy, nor by whether ρ_{Y} is small or large.

Comparing the results of the 2^{nd}, 3^{rd}, and 4^{th} rows in each panel (Figures 5.1 and 5.2), we see that the coverage rate is affected by the choice of SE estimation and also by whether SSDF correction is used. When the number of clusters is small, an SSDF correction must be used as failing to do so results in under-coverage (Panel A). The low coverage is more serious when TSLS adjusts for W (second and fourth set of results in each panel).

Overall, the results in Panel A of each figure show that using HW SE or not has little to no impact if there is no SSDF correction. However, when the SSDF correction is used for settings with cluster-level non-adherence, large ρ_{Y}, and large true LATE, but where only X is strongly associated with Y, using "unCL" outcomes leads to under-coverage, regardless of weighting or SE method (Figures 5.1 and 5.3, 3^{rd} and 5^{th} rows of Panel A, right hand side columns). The use of "adCL" outcomes (*i.e.* where the CL outcome is the residual after adjusting for individual level variable X) recovers coverage close to nominal. This is not the case when non-adherence is at the individual level, and both W and X are confounders of the causal effect of the treatment received D on the outcome Y in the data generating process.

In both cluster and individual-level non-adherence settings, it can be seen that using MV weights increases the coverage by a small fraction, when compared with cluster size weights, especially for scenarios with $J = 50$ and large ρ_{Y}. However, since MV weights require an estimate of the cluster-level variance, and this is badly estimated when the number of clusters is small ($J = 10$), we can see that MV weights are less efficient than using either no weights or cluster size weights. This is most clearly seen when no HW SE correction has been used.

We can also see that when SSDF correction is used, then not using HW SE can result in small over-coverage especially for cluster-level non-adherence settings, which is improved when HW SE are used (Figures 5.1 and 5.3, 3^{rd} and 5^{th} rows of Panel A).

When $J = 50$ (Panel B), the use of SSDF-based distributions is not expected to make any material difference, and this is indeed the case. The impact of using HW SE or the different weighting strategies is also minimal.

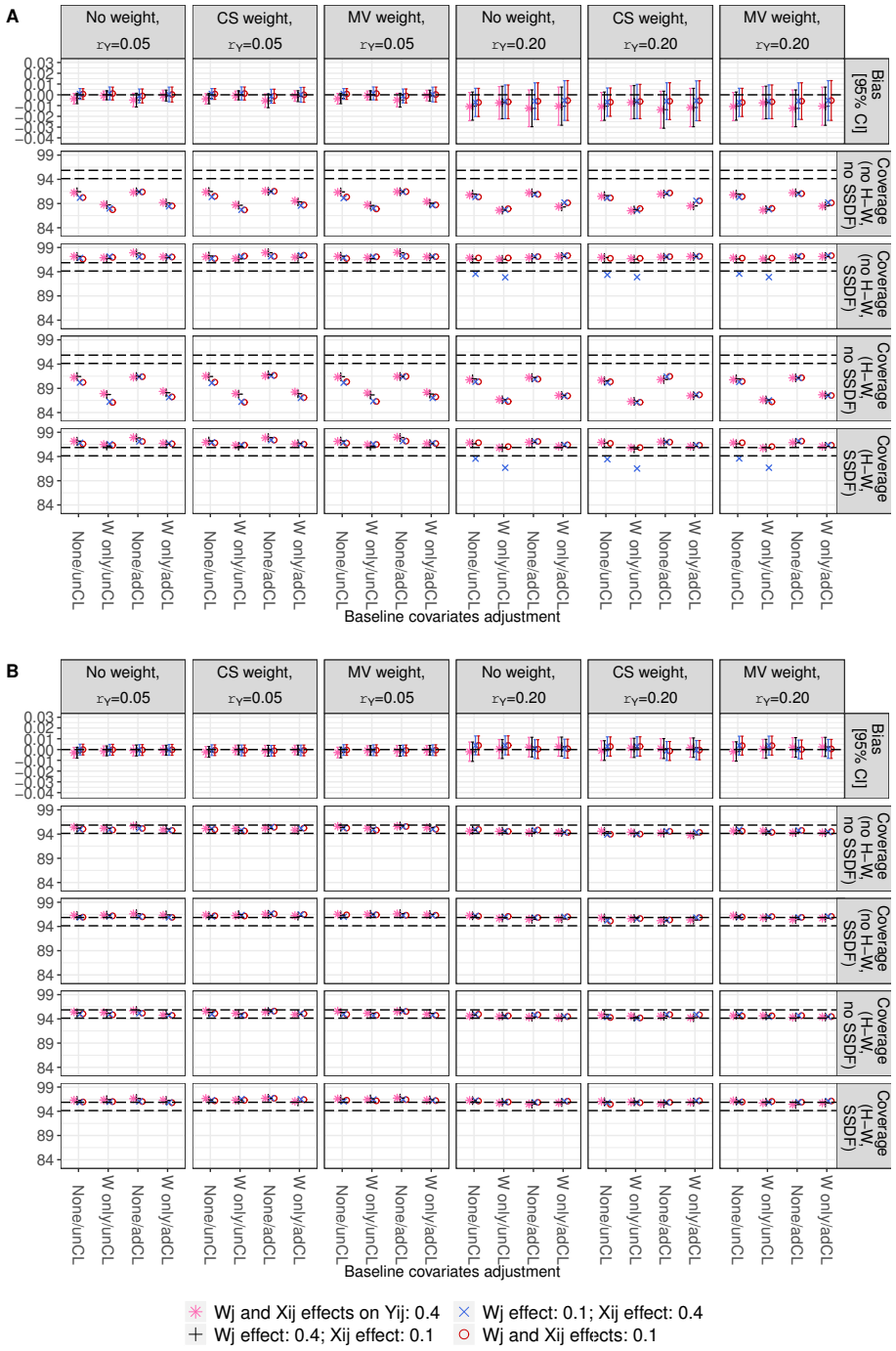


Figure 5.1: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with cluster-level non-adherence and modest true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B.

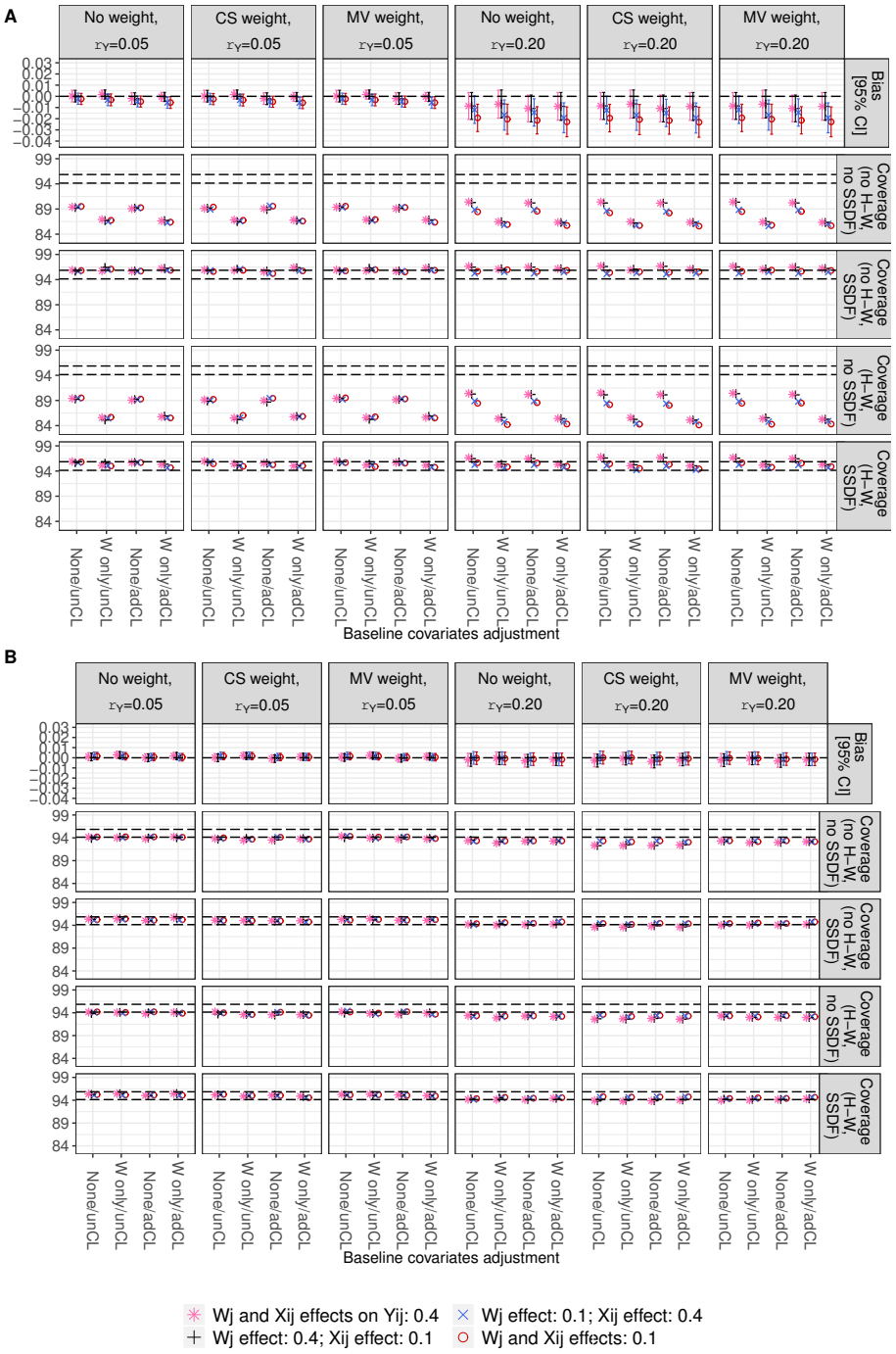


Figure 5.2: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and modest true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B.

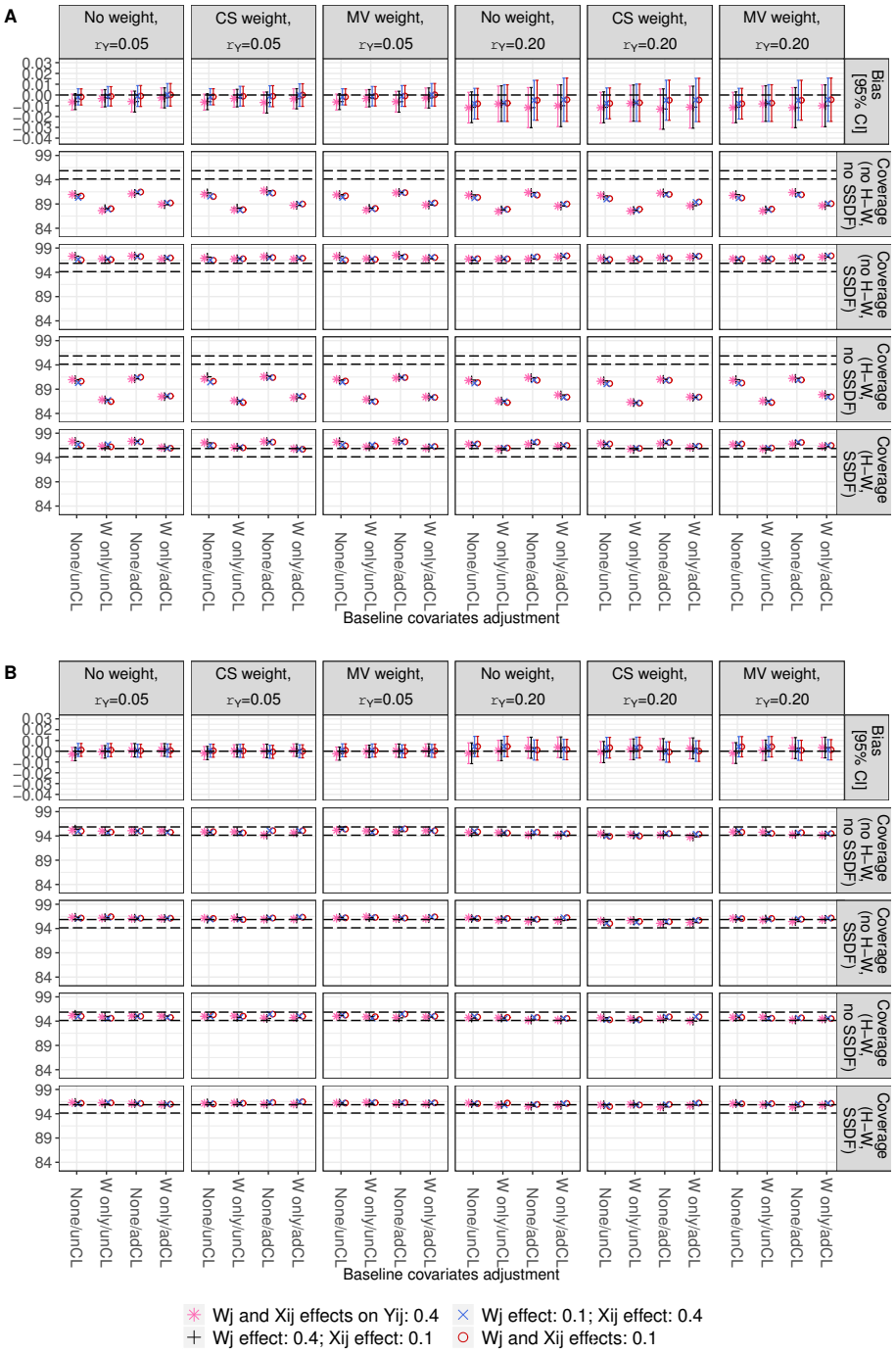


Figure 5.3: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with cluster-level non-adherence and small true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B.

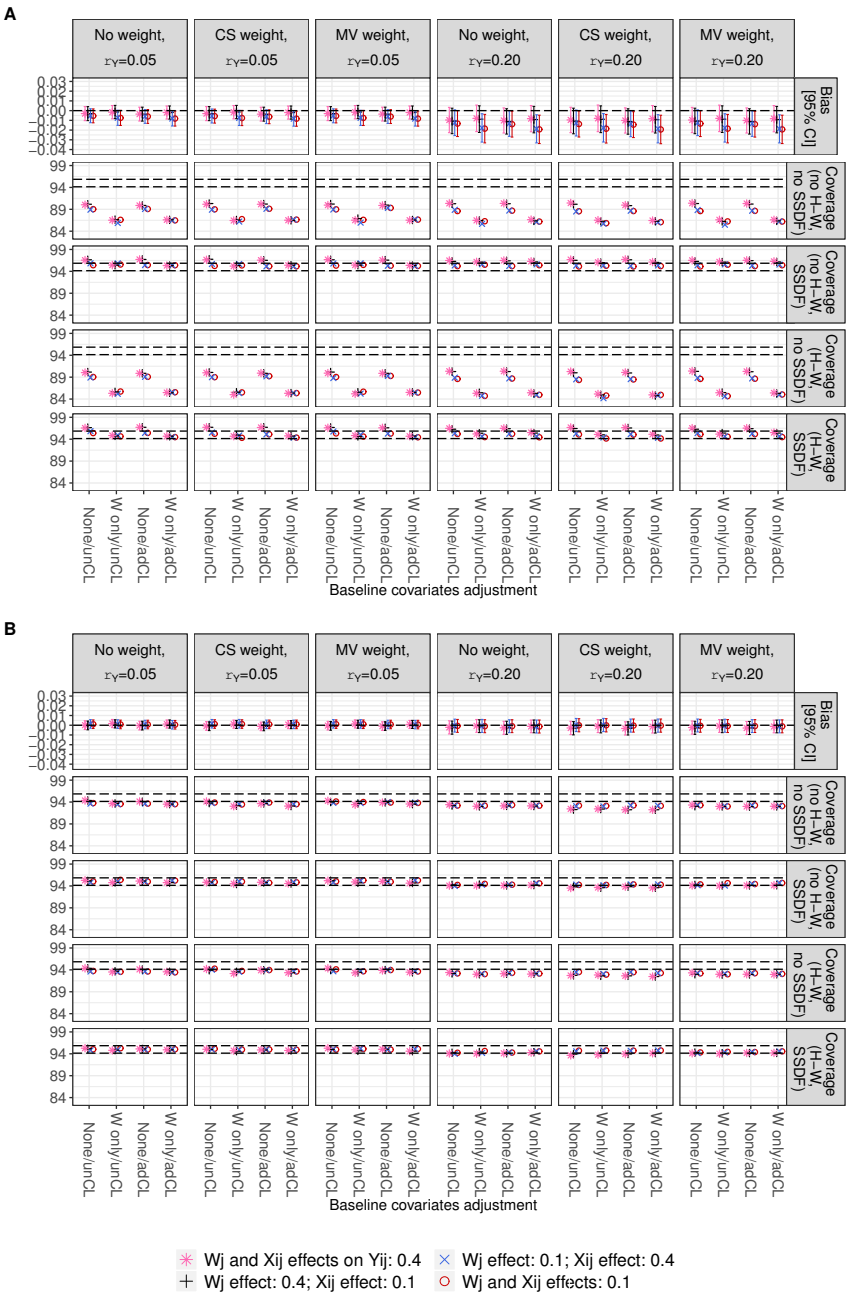


Figure 5.4: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and small true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B.

5.4.2 Schochet-Chiang approach

Figures 5.5 and 5.6 display the empirical bias and 95% CI coverage resulting from the Wald estimator with Schochet-Chiang SEs (all of which I refer to as the Schochet-Chiang estimator), when non-adherence is at the cluster and the individual levels respectively. Each Figure shows results for scenarios where the true LATE is small

(0.1 SD) and modest (0.4 SD) with varying ICC for the outcome (low or high). As above, each Figure also reports results where $J = 10$ (Panel A, top) or $J = 50$ (Panel B). The different data generation scenarios are identified by $*, +, \times$, and $\circ$, corresponding to varying strengths of the effects of X and W (considered as either observed or unobserved confounding) on Y. The scenarios considered here only include clusters with similar size.

The Schochet-Chiang estimator alike TSLS estimator shows finite sample bias irrespective of the level of adherence (whether at the cluster or at the individual level). Recall that the Schochet-Chiang method uses the Wald estimator where the numerator and denominator are estimated via OLS. In the absence of covariate adjustment, the Wald and TSLS estimators are equivalent [80]. The Schochet-Chiang's SEs are severely underestimated when the number of clusters is small and irrespective of the level of adherence (Panel A), leading to poor coverage. For settings with small number of clusters in particular, I note that when the ICC for Y is low, LATE size is large and CL covariate W is adjusted for while estimating LATE, the Schochet-Chiang's SEs result in values very close to 0 and sometimes are negative. Recall that the Schochet-Chiang's SEs are based on the between-variance which are small here as ICC for Y is low. Moreover, the between-cluster variance gets smaller when adjusting for W which, in my simulations, is associated with the outcome Y and the treatment received D. In contrast, the subtracted term in the Schochet-Chiang's formula of SEs (equation (4.28)) gets larger with increasing LATE size, making the variance of the Wald estimator to get close to 0 or negative. This is attenuated when the ICC is high as the between-cluster variance gets larger even though LATE size is large, leading to improved coverages.

The Schochet-Chiang estimator has good coverage when the number of clusters is large irrespective of the settings or when no covariate adjustment is done except for the settings with low ICC for Y and large LATE size. For those settings, the performance of the Schochet-Chiang and TSLS methods are comparable. However, the TSLS estimator with SSDF correction are preferable to the Schochet-Chiang estimator especially for settings where the number of clusters is small.

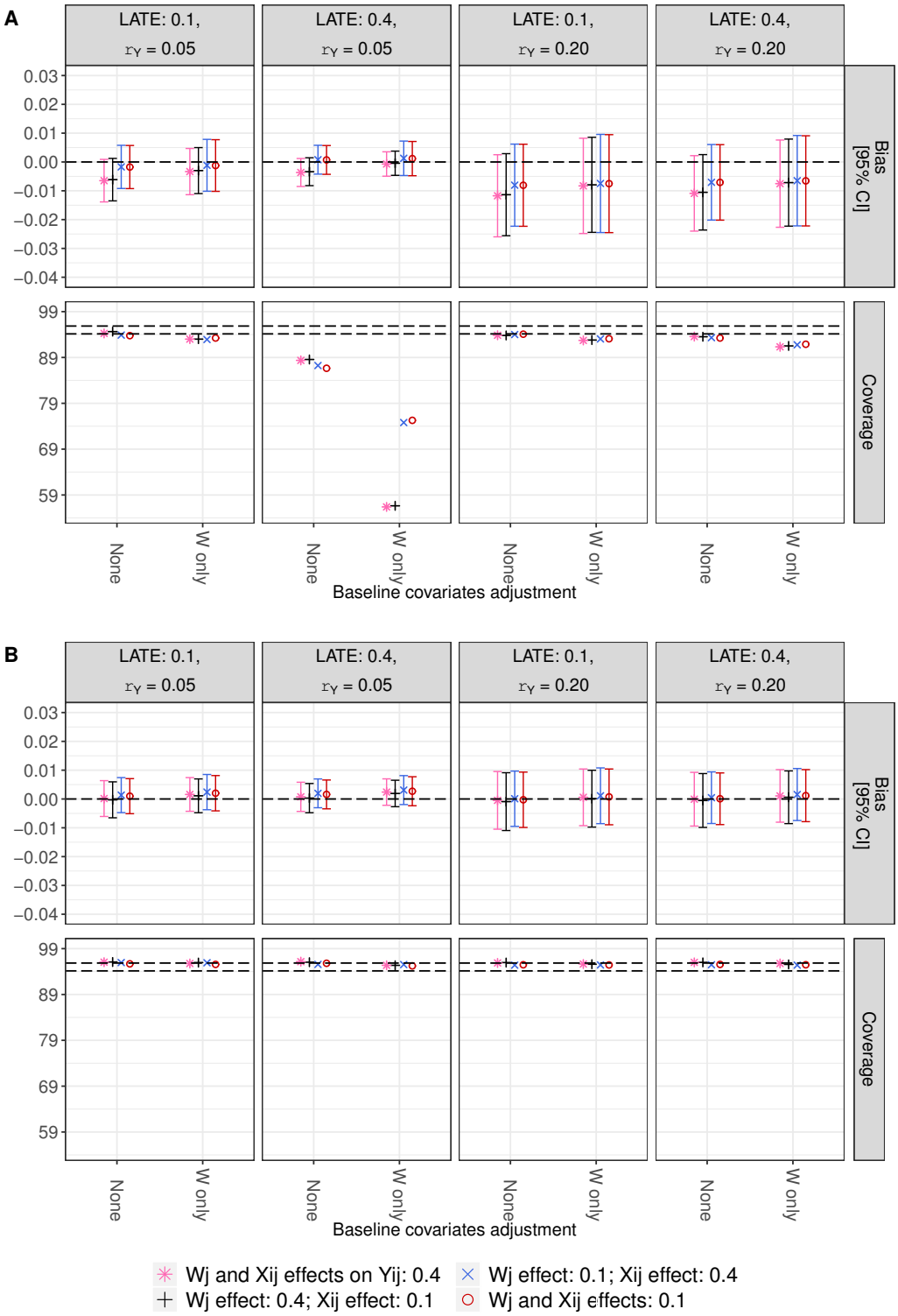


Figure 5.5: Bias (top row) and 95% CI coverage of CL-LATE with cluster-level non-adherence. The true LATE size and the ICC for outcome vary by columns. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained using the Wald estimator with Schochet-Chiang SEs without weighting and unadjusted or adjusted for W. Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B. The long-dashed black parallel lines are the acceptable 95% coverage range in the second panel.

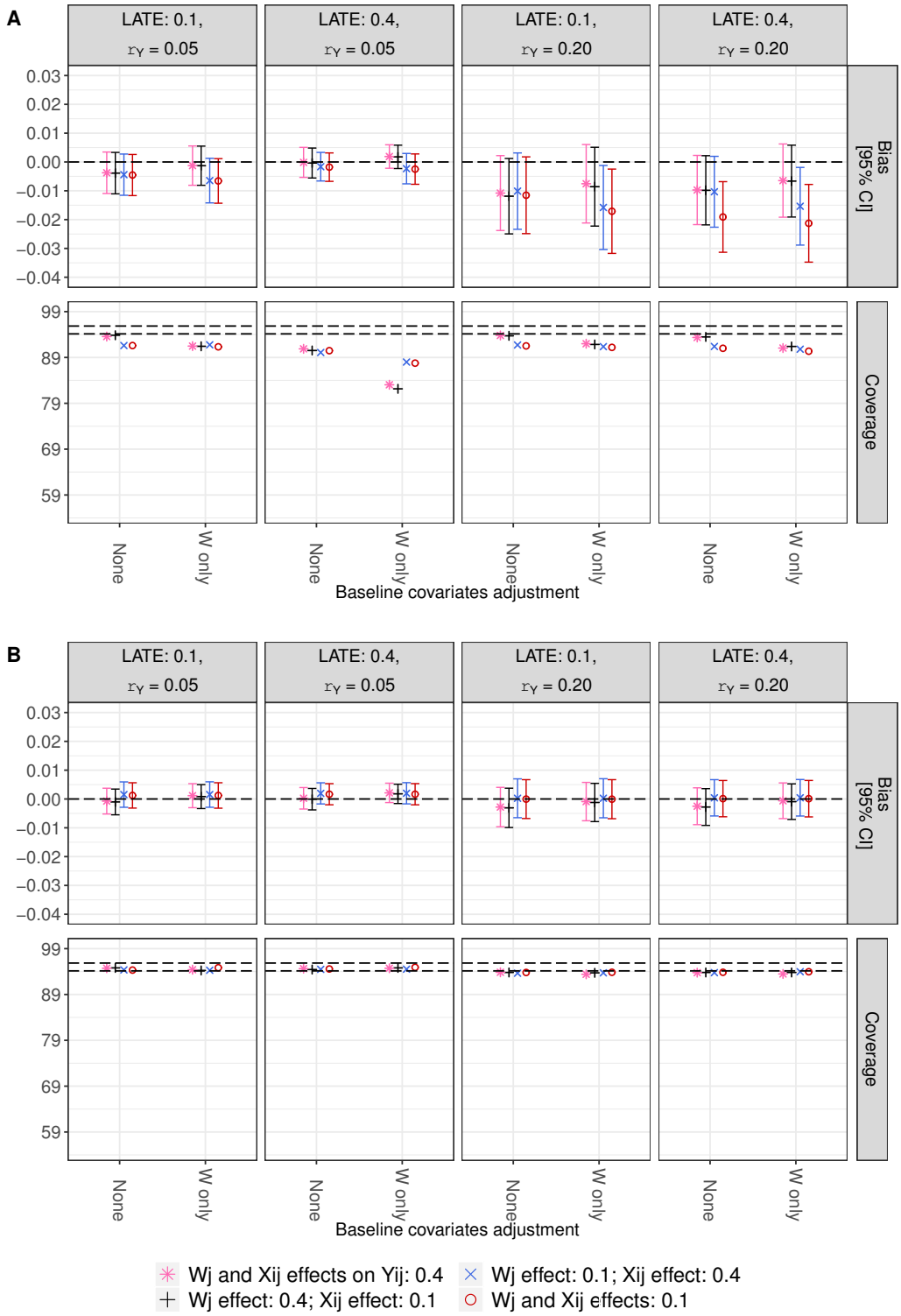


Figure 5.6: Bias (top row) and 95% CI coverage of CL-LATE with individual-level non-adherence. The true LATE size and the ICC for outcome vary by columns. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained using the Wald estimator with Schochet-Chiang SEs without weighting and unadjusted or adjusted for W. Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B. The long-dashed black parallel lines are the acceptable 95% coverage range in the second panel.

5.5 Additional simulations

Two extra additional scenarios are now considered to investigate the sensitivity of the performance of CL-TSLS and the Schochet-Chiang estimators' to number of clusters and cluster size imbalances, at both cluster and individual level adherence, but focusing on settings where confounding is strong with a modest true LATE.

In the first additional simulation, I explore the impact that the outcome ICC and the number of clusters have on bias, while leaving the expected total sample size fixed ($n = 1000$). We consider two marginal ICC for Y_{ij} ($\rho_Y = 0.05$ and $\rho_Y = 0.80$) and three average cluster size ($n_j = 20$, 10 and 2.5, corresponding to whether the number of clusters varied from $J = 50, 100$ or 400), which includes one of the scenarios previously considered in the main simulations for comparison. Though CRTs rarely have ICCs above 0.10 [107], the value of $\rho_Y = 0.80$ is included to evaluate the performance of the methods in extreme settings.

In the second additional set of simulations, I explore the effect of high cluster size imbalances. While keeping the average sample size equal to 1000, and $J = 10$ or 50, I create high cluster size imbalance using a Pareto distribution to generate the cluster sizes [108]. The Pareto distribution parameters are chosen so that approximately 40% of the clusters have a size below 15, and 60% a size above 15, while the average cluster size is 20 and the minimum cluster size is 10, resulting in approximately 1.8 for the shape and 9.1 for the scale.

5.5.1 Results from TSLS estimation

Figures 5.7 and 5.8, corresponding to cluster and individual level non-adherence settings, show that for a fixed number of clusters (cells in the same row), the bias increases with increasing ICC for Y, but that as the number of clusters increase (moving down the column in the Figure), CL-TSLS results in negligible mean bias, even for a very large ρ_Y. It is well known that TSLS is only asymptotically unbiased, and with CL analyses, we expect the asymptotics to depend on the number of clusters, and not the number of individuals. Nevertheless, the CL-summaries treated as outcomes for the two models involved in TSLS contain less "information" when

the ICC is higher, which translates into a larger number of clusters being necessary for the bias to be negligible.

The impact of high cluster size imbalance is reported in Figures 5.9 and 5.10, where non-adherence is at the cluster and individual level respectively. We see that when $J = 10$ (Panel A), even with SSDF correction, failure to use HW SE results in under-coverage when using cluster size weights, which is especially pronounced when ρ_Y is large. This is because cluster size weights are known to perform well when the cluster level residuals are homoscedastic, which is unlikely when cluster sizes are very imbalanced [76]. This also explains why using HW SE brings the coverage close to nominal levels. This pattern is also observed at $J = 50$ (Panel B).

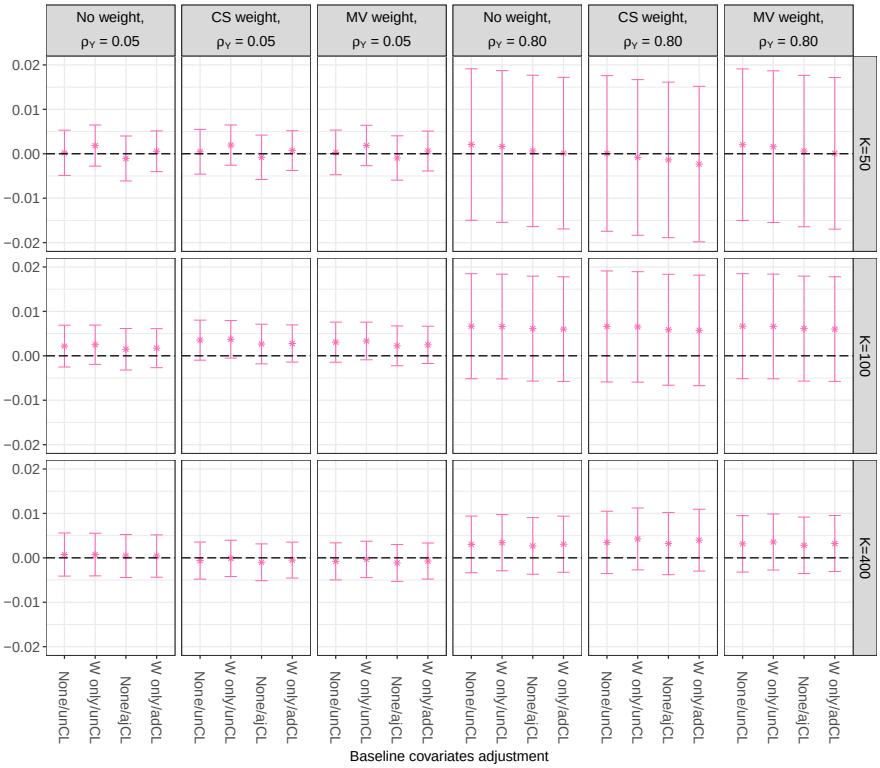


Figure 5.7: Bias of the CL-LATE for the extra simulation where non-adherence is at the cluster level and a modest true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

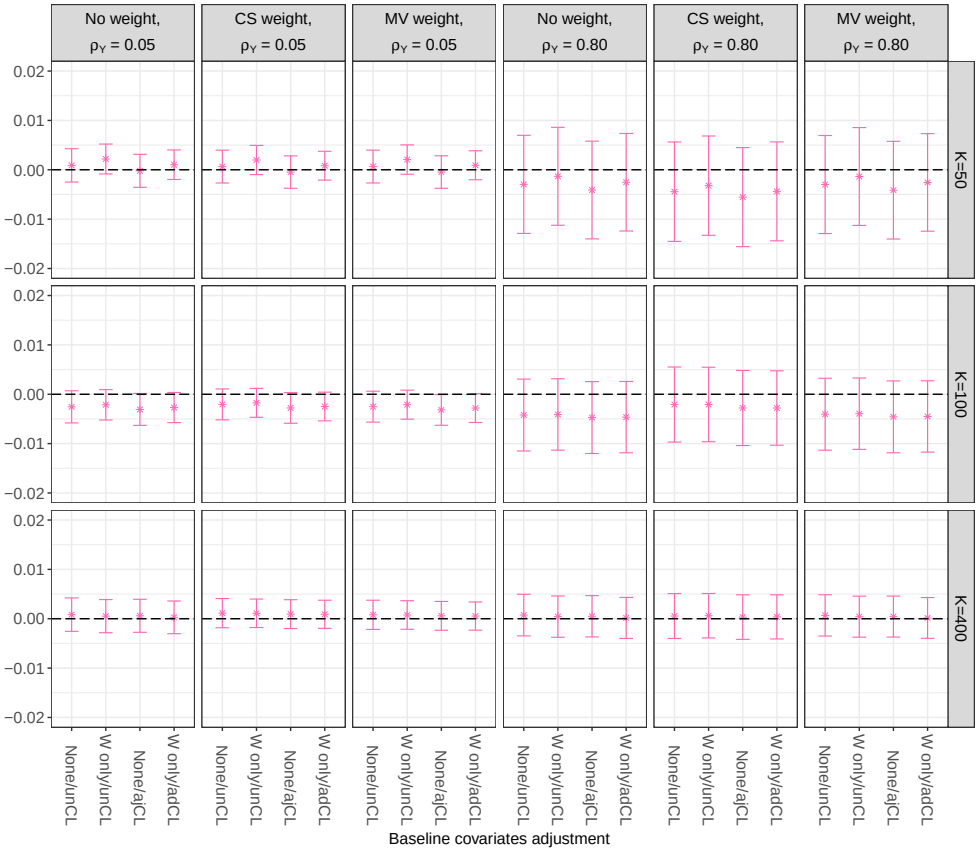


Figure 5.8: Bias of the CL-LATE for the extra simulation where non-adherence is at the individual level and a modest true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

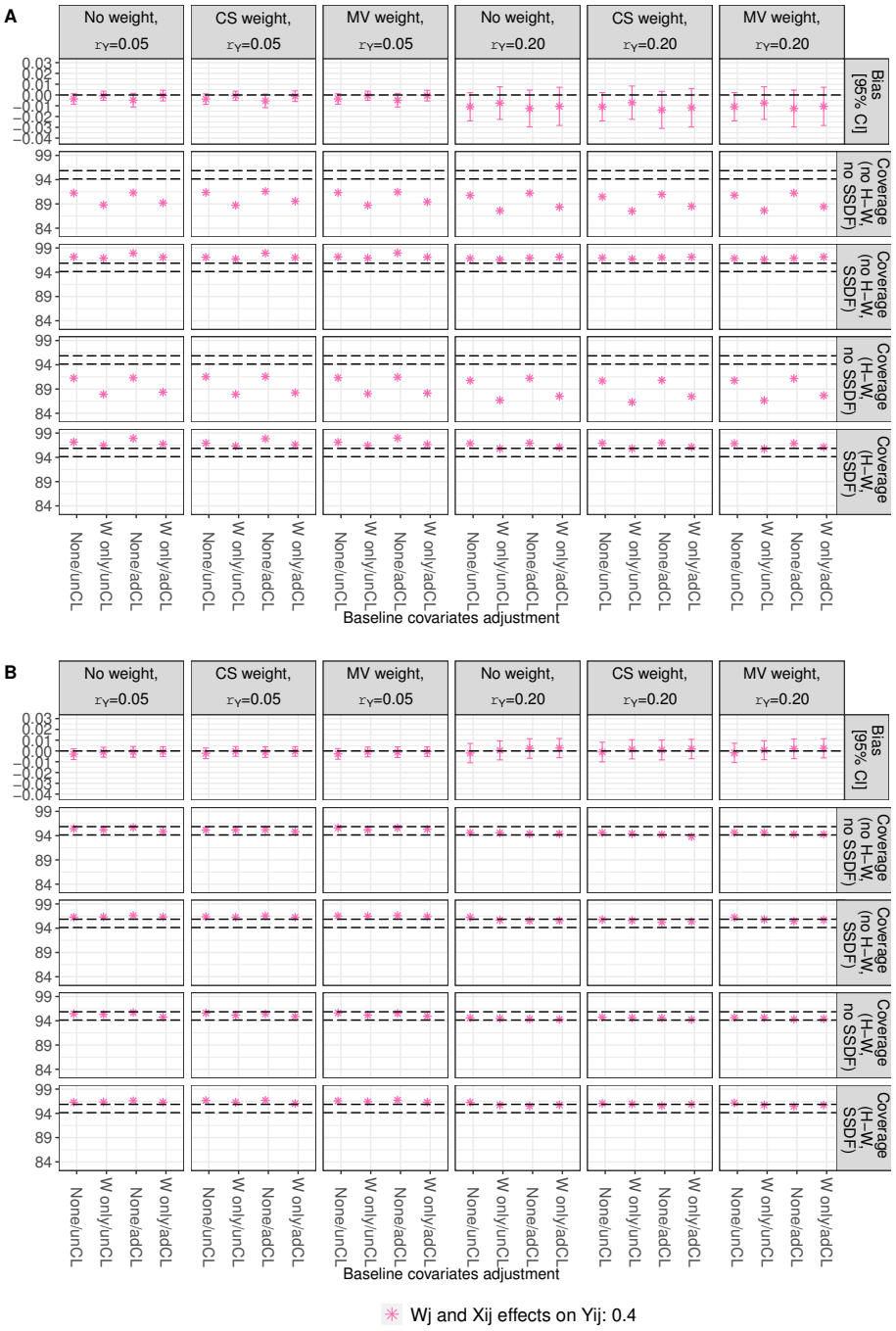


Figure 5.9: Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber-White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the cluster level, and a modest true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Small and large number of clusters results appear in Panels A and B respectively.

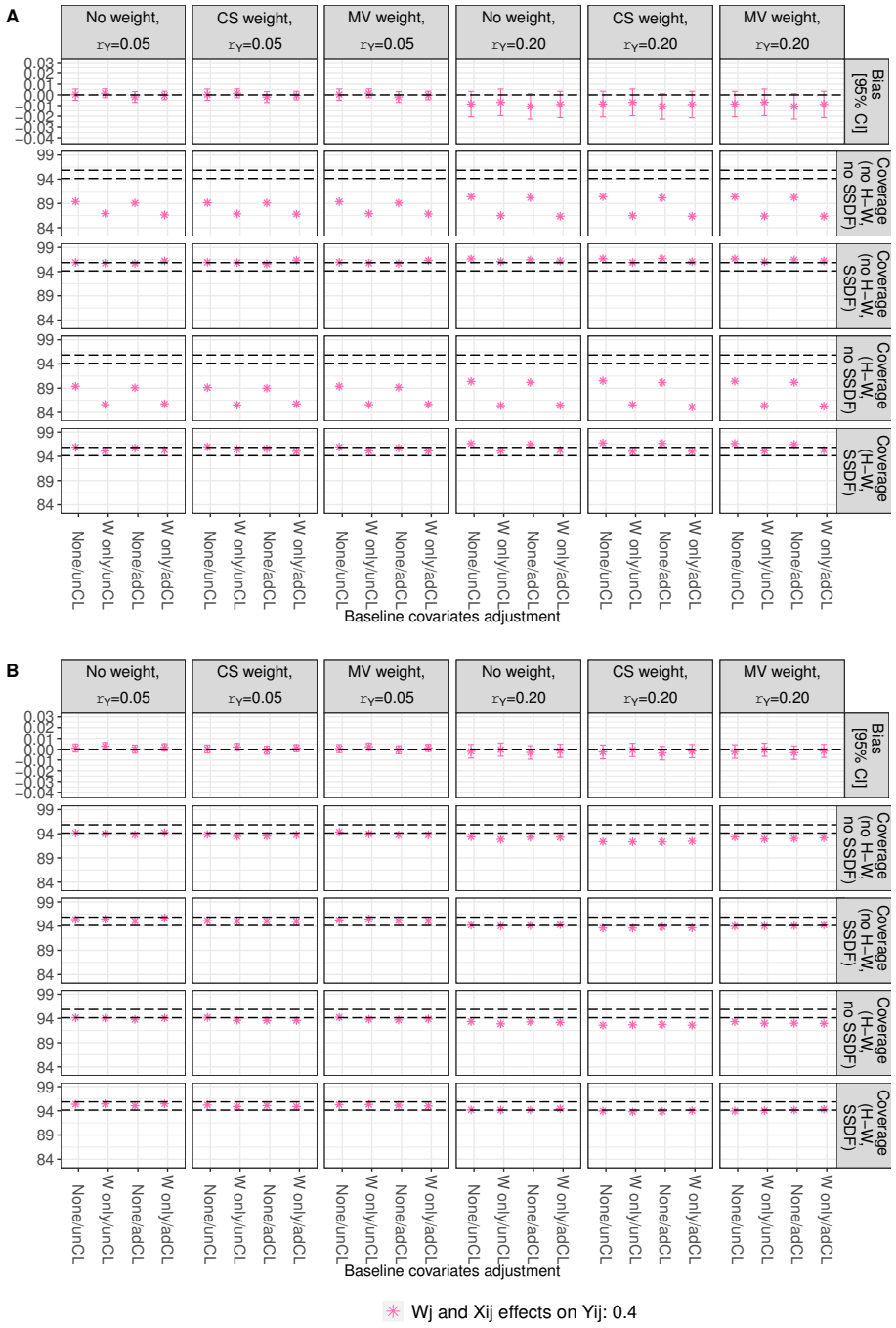


Figure 5.10: Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber-White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the individual level, and a modest true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Small and large number of clusters results appear in Panels A and B respectively.

5.5.2 Results from Schochet-Chiang approach

Figures 5.11 displays the bias for cluster-level (panel A) and individual-level (panel B) non-adherence settings for increasing number of clusters. Like TSLS, for a fixed

number of clusters (cells in the same row), the bias exacerbates with increasing ICC for the outcome. When the ICC for Y is large, the bias decreases as the number of clusters increases. However, there is little impact of increasing the number of clusters on the bias when the ICC or outcome is low. This is not surprising because the Schochet-Chiang estimator uses information at the cluster level and thus with high ICC or Y, the between-cluster variance is larger but increasing the sample size (*i.e.* the number of clusters) is necessary to improve the efficiency of the Schochet-Chiang estimator. The larger the ICC for Y is, the greater should the number of clusters be to attenuate the bias.

The impact of high cluster size imbalance is shown in Figure 5.12, where non-adherence is at the cluster level (Panels A and B) and at the individual level (Panels C and D). Results from scenarios with small number of clusters are presented in Panels A and C whereas those with large number of clusters are shown in Panel B and D. Like TSLS, irrespective of the number of clusters, the cluster size imbalance has little effect on the performance of the Schochet-Chiang estimator in terms of empirical bias. Moreover, the Schochet-Chiang estimator shows good coverage irrespective of the scenarios. The Schochet-Chiang's performance is similar to that of TSLS with SSDF adjustment and HW SE, whether non-adherence is at the cluster or individual level.

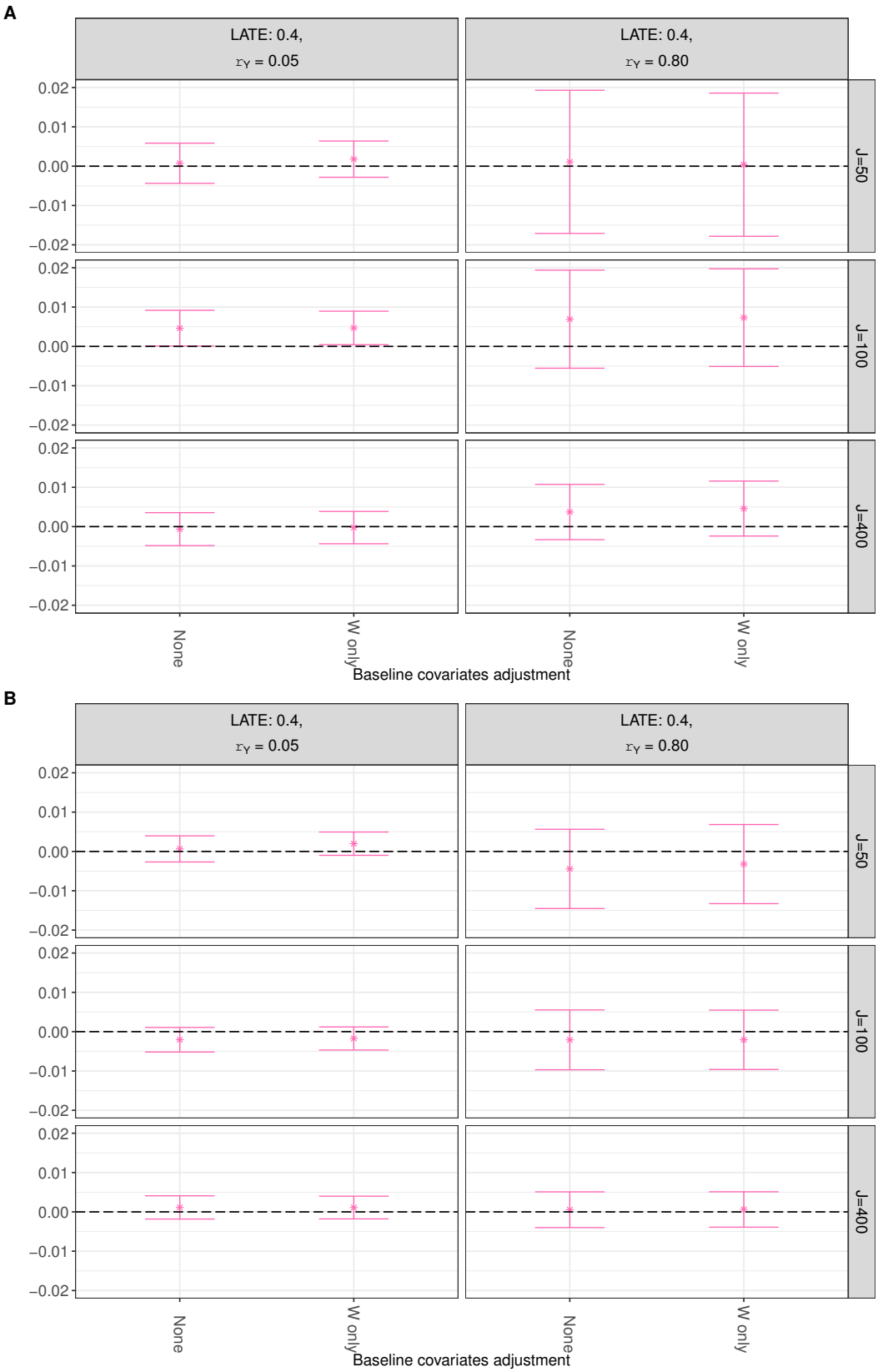


Figure 5.11: Bias of the CL-LATE for the extra simulation where non-adherence is at the cluster level (Panel A) and at the individual level (Panel B). The true LATE size is modest, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted Schochet-Chiang method without weighting. Number of clusters varies by rows and ICC by column.

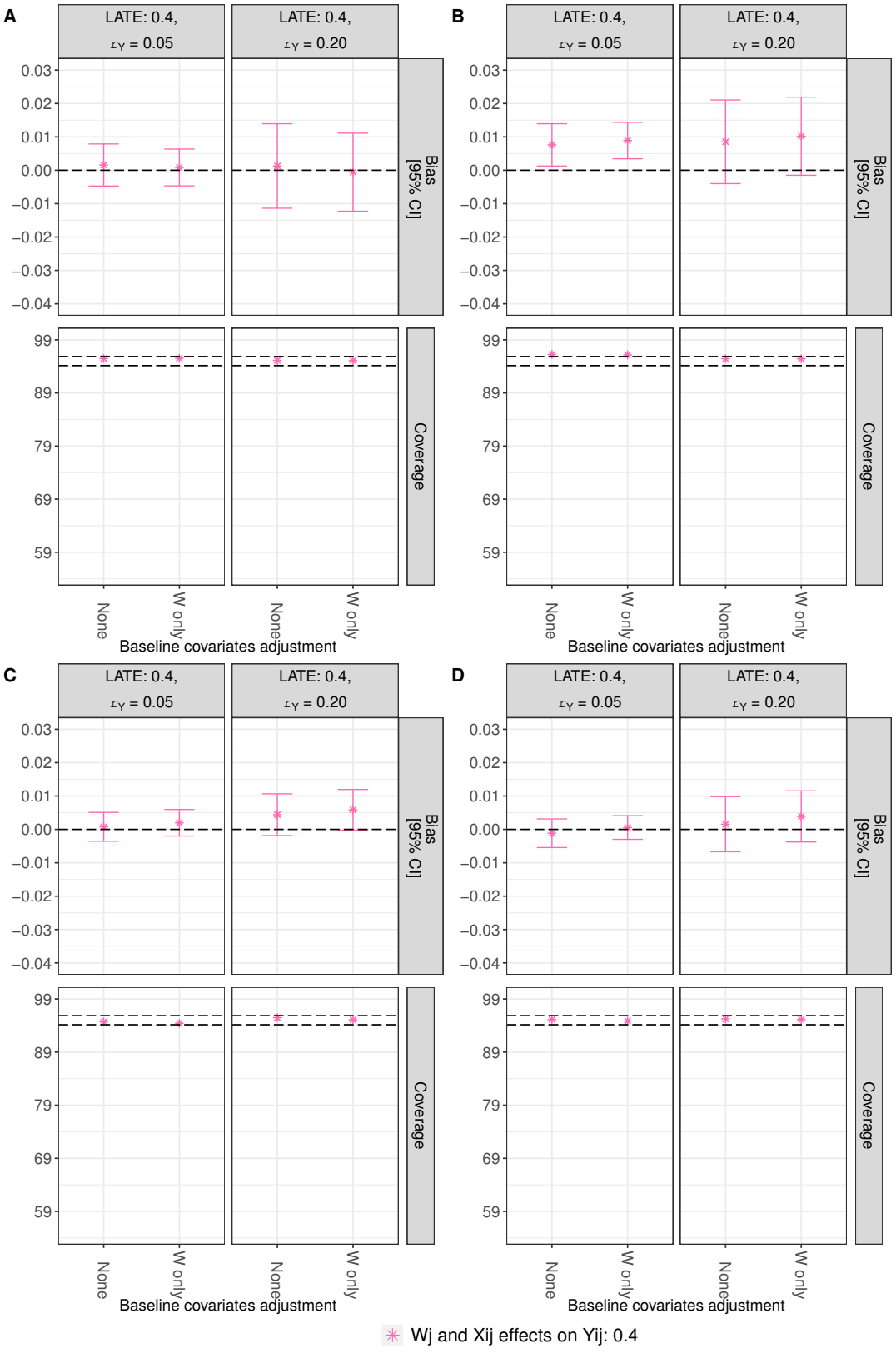


Figure 5.12: Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage of the CL-LATE using Schochet-Chiang method without weighting, where non-adherence is at the cluster level (Panels A and B) and at the individual level (Panels C and D). Small number of clusters results appear in Panels A and C and large number of clusters results in Panel B and D. The true LATE size is modest. The long-dashed black parallel lines are the acceptable 95% coverage range in the second panel.

5.6 Summary

I investigate the performance of Schochet-Chiang approach and TSLS to estimate CL-LATE in terms of empirical bias and coverage through simulation study. For TSLS estimation, different weighting strategies (none, cluster size, minimum variance) and methods for obtaining CIs (alternatively using or not HW SEs and/or SSDF correction) have been explored. TSLS estimator is known in general to result in finite sample bias but no simulation study has focused on TSLS performance nor on the Schochet-Chiang approach as an alternative to TSLS in CRTs when the causal treatment effect at the cluster level is of interest. I explore settings for one-sided non-adherence CRTs of different sizes and where non-adherence is either at the cluster or individual level, and I allow for various effect sizes of cluster-level and individual-level variables on the outcome and the treatment received. I only simulated CRTs where the random treatment assignment is a relevant instrument. This inclusion criterion is to reflect well designed and conducted trials where there are no major issues with the acceptability of the intervention. I demonstrate the use of TSLS applied to CL summaries as a simple and valid method for obtaining estimates of the LATE in CRTs where non-adherence occurs at the cluster or the individual level.

Empirically via simulations, under the sufficient assumptions for identification, TSLS regression of CL summaries provides consistent estimates of the causal treatment effect in the sub-population of compliers, where non-adherence is at the cluster level. With individual-level non-adherence, the additional assumption that the cluster-specific LATE is homogeneous across clusters is required for CL-TSLS to identify the population LATE [98]. Moreover, provided that an appropriate distribution with SSDF adjustment is used when the number of clusters is small and HW SEs are used if there is high cluster size imbalance, valid 95% CIs can be constructed.

The simulations suggest that all weighting strategies perform similarly when the number of clusters is not small. When the number of clusters is small, MV weights tend to be badly estimated and are not recommended. Furthermore, when the

cluster sizes are very variable, CS weights should not be used. Although in the simulations the weights did not affect the point estimates, these were affected in the illustrative example (see chapter 8). Overall the results show that unless there are very few clusters, or the outcome ICC is large, MV weighting performs well [74].

The Schochet-Chiang estimator has good coverage when the number of clusters is large irrespective of the settings or when no covariate adjustment is done except for the settings with low ICC for Y and modest LATE size. For those settings, the performance of the Schochet-Chiang and TSLS methods are comparable. However, the TSLS estimator with at least SSDF correction is preferable to the Schochet-Chiang estimator especially for settings where the number of clusters is small.

Chapter 6.

Estimation of local average treatment effect at individual level in CRTs

6.1 Introduction

Data from CRTs are often analysed at the individual level. The systematic review presented in chapter 2 shows that about 97% of the reviewed CRT reports used IL analyses, of which 70% are performed via generalized estimating equations or mixed effects modelling (Table 2.2). IL analysis has the advantage over CL summary approach (introduced chapter 3) to easily adjust for both baseline CL and IL covariates.

The present chapter is structured as follows. In section 6.2, I introduce the assumptions required for the identification of IL-LATE. Section 6.3 presents some IL-LATE estimation options such as the Wald estimator with bootstrapped SEs, TSLS with cluster robust SEs and Moulton's corrected SEs, multilevel mixture modelling using a Bayesian approach. I also provide an overview of multilevel mixture modelling via the expectation-maximization (EM) algorithm, which can be implemented in Mplus [109]. Those methods are implemented in chapters 7 and 9, except the EM analysis due to inaccessibility to Mplus. Though not covered in this thesis, note that TSLS could also be used along with bootstrapped SEs. However, I focus on TSLS with cluster robust SEs which is popular for causal inference using clustered data and TSLS with Moulton's corrected SEs which is attractive and simple but not often used in practice.

6.2 Identification of IL-LATE

I formally introduce the causal estimand LATE at the individual level, and the assumptions for its identification within the POs framework as applied to CRTs in the presence of binary treatment assignment and binary non-adherence.

6.2.1 Notation and technical assumptions

Like in chapter 4, I consider a two-arm CRT, with n individual units indexed by i, and J clusters, labelled by j, each of size n_j. Let Z_j be the binary random treatment assignment for cluster j, Y_{ij} be the continuous, and $D_{ij} \in \{0, 1\}$ be the treatment received by individual i in cluster j. Let W_j and X_{ij} be CL and IL baseline covariates, respectively (which can be vectors of variables) and associated with both Y_{ij} and D_{ij}. Assume that there exists at least one unobserved confounder of the D_{ij}–Y_{ij} relationship, denoted by U_{ij} .

The same notation and technical assumptions mentioned in section 4.2.1 are also needed for IL-LATE settings *i.e.* assuming *no interference between clusters* and *counterfactual consistency.*

6.2.2 IL-LATE estimand

Let C_{ij} denote the adherence class [27, 34] for individual i in cluster j: $C_{ij} = n$ (never-takers) if $D_{ij}(0) = D_{ij}(1) = 0$; $C_{ij} = a$ (always-takers) if $D_{ij}(0) = D_{ij}(1) = 1$; $C_{ij} = c$ (compliers) if $D_{ij}(z) = z$ for $z \in \{0, 1\}$; and $C_{ij} = de$ (defiers) if $D_{ij}(z) = 1 - z$ for $z \in \{0, 1\}$. Recall that the non-numerical values n, a, c and d are used to simply help recognising the *adherence classes.* IL-LATE is the same as the *population* LATE defined in equation (4.1) shown in section 4.2.2, that is,

$$\begin{aligned}
\beta &= \mathbb{E}\Big[\big\{Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\big\}\Big|C_{ij} = c\Big] \\
&= \mathbb{E}\Big[Y_{ij}(1) - Y_{ij}(0)\Big|D_{ij}(1) - D_{ij}(0) = 1\Big] \\
&= \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\Big[Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\Big]\Big[I\Big(D_{ij}(1) = 1, D_{ij}(0) = 0\Big)\Big]}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} I\Big(D_{ij}(1) = 1, D_{ij}(0) = 0\Big)}
\end{aligned}$$

It is also necessary that assumptions **(A1)-(A4)** introduced in section 4.2.2 hold to identify β. Under the relevant technical (*no interference between clusters* and

counterfactual consistency) and identification (IV and *monotonicity*) assumptions, the causal estimand β can be simplified using the observed values rather that the *POs* as follows

$$\beta_{\text{IL}} = \frac{\mathbb{E}\big[Y_{ij}|Z_j = 1\big] - \mathbb{E}\big[Y_{ij}|Z_j = 0\big]}{\mathbb{E}\big[D_{ij}|Z_j = 1\big] - \mathbb{E}\big[D_{ij}|Z_j = 0\big]}$$

β_{IL} is the ratio of the ITT effect on Y_{ij} (causal effect of Z_j on Y_{ij}) to the ITT effect on D_{ij} (causal effect of Z_j on D_{ij}). When IL-LATE is adjusted for a single covariate W_j for instance (assumed here to be categorical), the adjusted IL-LATE is formulated as

$$\beta_{\text{IL,adj}} = \frac{\sum_{j=1}^{J}\Big(\mathbb{E}\big[Y_{ij}|Z_j = 1, W_j\big] - \mathbb{E}\big[Y_{ij}|Z_j = 0, W_j\big]\Big)\Pr(W_j)}{\sum_{j=1}^{J}\Big(\mathbb{E}\big[D_{ij}|Z_j = 1, W_j\big] - \mathbb{E}\big[D_{ij}|Z_j = 0, W_j\big]\Big)\Pr(W_j)}$$

where $\Pr(W_j)$ is the empirical distribution of the observed Wj.

Note that the clustering does not affect the expected values. However, clustering must be accounted for when estimating the variance of β_{IL}'s estimates.

6.3 Estimation of IL-LATE

We introduce two popular approaches for estimating β, which are the Wald and TSLS estimators and the mixture modelling. The former approach is the traditional IV estimation and the latter explicitly models both the latent adherence class and the outcome.

6.3.1 TSLS estimation

As introduced in chapter 4, TSLS provides a consistent estimate for β [76], which I denote here $\widehat{\beta}_{\text{IL,TSLS}}$. The first and second stages of TSLS estimation are OLS regressions (ignoring clustering) as follows, considering that covariates are included:

first stage,

$$D_{ij} = \gamma_0 + \gamma_{\text{Z}} Z_j + \gamma_{\text{W}} W_{ij} + \gamma_{\text{X}} X_{ij} + \epsilon_{1_{ij}} \tag{6.1}$$

and second stage,

$$Y_{ij} = \beta_0 + \beta_{\text{IL,TSLS}} \widehat{D}_{ij} + \beta_{\text{W}} W_j + \beta_{\text{X}} X_{ij} + \epsilon_{2_{ij}} \tag{6.2}$$

where $\widehat{D}_{ij}$ is the predicted value of D_{ij} from the first stage, $\epsilon_{1_{ij}} \sim N(0, \sigma^2_{\epsilon_1})$ and

$\epsilon_{2_{ij}} \sim N(0, \sigma^2_{\epsilon_2})$. $\widehat{D}_{ij}$ are afterwards used as a covariate as described in equation 6.2. I present below two approaches for estimating the standard error of $\widehat{\beta}_{\text{IL,TSLS}}$, acknowledging the clustering. These approaches are the Huber-White-Rogers (also known as cluster-robust) and Moulton's standard errors.

6.3.1.1 Huber-White-Rogers standard error

When data are clustered, ignoring the correlation between units within clusters may bias the standard errors downward. However, a simple way of estimating β is to implement TSLS estimation ignoring the clustering and subsequently correct the standard error using Huber-White-Rogers method [70, 110]. Huber-White standard errors known also as robust or sandwich standard errors are consistent even in the presence of heteroscedastic residuals provided that the residuals are independently distributed. The Huber-White-Rogers method is an extension to the Huber-White approach that relaxes the assumption of independent residuals and produces consistent standard errors in the presence of clustering, provided that individual units are correlated within clusters but independent across clusters. The point estimates remain unchanged.

6.3.1.2 Moulton standard error correction

This estimation procedure is based on equations 6.1 and 6.2 as above, but the conventional OLS standard error of $\hat{\beta}_{TSLS}$ is inflated by a scalar referred to as Moulton factor [76]. The Moulton factor for $\hat{\beta}_{TSLS}$ is the ratio of $\hat{\beta}_{TSLS}$'s standard error obtained from fitting linear random-intercept regressions (adding random intercepts in equations 6.1 and 6.2 to account for the clustering) to its conventional standard error from OLS regressions (as in equations 6.1 and 6.2). The Moulton factor is originally applied to correct standard errors for valid inference posterior to the use of linear regressions when analysing hierarchical data, but is extended to TSLS estimation [76, 111] as follows:

$$\frac{\text{SE}(\hat{\beta}_{TSLS})}{\text{SE}_c(\hat{\beta}_{TSLS})} = \Big(1 + \Big[\frac{\text{Var}(n_j)}{\bar{n}} + \bar{n} - 1\Big]\rho_{\bar{D}}\rho_{\epsilon_2}\Big)^{\frac{1}{2}} \tag{6.3}$$

where SE_c is the conventional standard error from equation 6.2 and SE the valid standard error accounting for clustering. $\text{Var}(n_j)$ is the variance of clusters size and

$\hat{n}$ the average cluster size. $\rho_{\hat{D}}$ and ρ_{ϵ_2} are the intra-cluster correlation coefficients of $\hat{D}_{ij}$ (predicted values of the first-stage) and $\epsilon_{2_{ij}}$ (residuals of the second stage), respectively. From the following random-intercept models $\hat{D}_{ij} = \alpha_0 + u_{1j} + \psi_{1_{ij}}$ and $\epsilon_{2_{ij}} = u_{2j} + \psi_{2_{ij}}$ where $u_{1j} \sim N(0, \sigma^2_{u_1})$, $\psi_{1_{ij}} \sim N(0, \sigma^2_{\psi_1})$ and $\psi_{2_{ij}} \sim N(0, \sigma^2_{\psi_2})$, we get $\rho_{\hat{D}} = \frac{\sigma^2_{u_1}}{\sigma^2_{u_1}+\sigma^2_{\psi_1}}$ and $\rho_{\epsilon_2} = \frac{\sigma^2_{u_2}}{\sigma^2_{u_2}+\sigma^2_{\psi_2}}$.

From equation 6.3, we note that the Moulton factor is not required when there is no clustering *i.e.* $\rho_{\hat{D}} = 0$ and/or $\rho_{\epsilon_2} = 0$; for instance, when the treatment received is at cluster level, ignoring clustering would give valid inferences. The higher the cluster size imbalance, the greater the downward bias of the standard error when clustering is present.

6.3.2 Wald estimator

The basis of the Wald estimator [31], denoted by $\widehat{\beta}_{\text{IL,Wald}}$, in equation (6.1) where Z_j is a binary IV and there is no covariate adjustment. $\widehat{\beta}_{\text{IL,Wald}}$ is a simple and consistent estimator of β [102]. Let us consider the following OLS regression models

$$Y_{ij} = \beta_0 + \beta_Z Z_j + \epsilon_{1_{ij}} \tag{6.4}$$

$$D_{ij} = \gamma_0 + \gamma_Z Z_j + \epsilon_{2_{ij}} \tag{6.5}$$

where $\epsilon_{1_{ij}}$ and $\epsilon_{2_{ij}}$ are *i.i.d* normal residuals with mean 0. The Wald estimator of β is

$$\widehat{\beta}_{\text{IL,Wald}} = \frac{\widehat{\beta}_{\text{Z}}}{\widehat{\gamma}_{\text{Z}}} \tag{6.6}$$

where $\widehat{\beta}_{\text{Z}}$ is the ITT effect on outcome (as the sample-means difference in Y_{ij} between control and active groups) and $\widehat{\gamma}_{\text{Z}}$ the ITT effect on treatment received.

Although the Wald estimator refers to the setting with single binary IV and without covariates, with a slight abuse of terminology, I label this ratio as the "adjusted" Wald estimator when covariates W_j and/or X_{ij} are included in equations (6.4) and (6.5). Equation (6.6) is analogous to the so-called "indirect least squares" estimator, where models (6.4) and (6.5) are fitted via OLS [76], *i.e.* ignoring clustering.

$\widehat{\beta}_{\text{IL,Wald}}$ is a non-linear combination of two estimators and therefore obtaining an analytic form of its variance may entail some approximations. Moreover for valid

inference, it is necessary to account for clustering in estimating the standard error of $\widehat{\beta}_{\text{IL,Wald}}$. The standard error of $\widehat{\beta}_{\text{IL,Wald}}$ can be estimated using the traditional Delta method or a bootstrap approach. The former is commonly obtained using the first order approximation of Taylor series expansion, known as the "Delta" method [103] (see section 4.4.2). For the inference based on bootstrapping, both clusters and individuals within clusters are re-sampled N times [80], and 95% CIs constructed based on normal approximation. Bootstrap CIs are suitable for small and large data sets and may avoid misleading inferences [112, 113]. As regards to how large N should be, it has been suggested that N should be between 1000 and 2000 [113].

6.3.3 Multilevel mixture model

The LATE can also be estimated by fitting an appropriate mixture model, with the LATE corresponding to a specific model parameter. Estimation of these models can be achieved within a frequentist or a Bayesian framework. Here I focus on Bayesian and the expectation-maximization (EM) estimations. These modelling techniques are based on two main specifications: the outcome model and the model for adherence. Clustering is accounted for by including a multilevel specification of the error structure, leading to multilevel mixture models being fitted to estimate the LATE.

In the setting of CRTs where there is one-sided non-adherence at the individual level, that is where there are only *compliers* and *never-takers* by design, the multilevel mixture model may be written as

$$Y_{ijl} \sim f(Y_{ij}|\theta_l, C_{ijl}, Z_j, W_j, X_{ij}), \qquad l \in \{1, 2\}, \qquad C_{ijl} \sim \text{Bern}(\pi_{ijl}) \tag{6.7}$$

where $f(Y_{ij}|\theta_l, C_{ijl}, Z_j, W_j, X_{ij})$ is the density probability of Y with parameter $\theta_l = (\boldsymbol{\beta_l}, \boldsymbol{\upsilon_l}, \sigma_l^2)$ where $\boldsymbol{\beta_l}$ represents the vector of fixed effects (regression coefficients) to be estimated, $\boldsymbol{\upsilon_l}$ is the vector of random effects for adherence class l. Note that each individual unit may have a non-zero probability $\pi_{ijl} = p(C_{ij} = l)$ to belong to adherence class l. We label *never-takers* as $l = 1$ and *compliers* as $l = 2$. Thus, the probability to be a *complier* is $\pi_{ij_2} = p(C_{ij} = 2) = \pi_{ij}$ and the probability of being

a *never-taker* is $\pi_{ij_1} = p(C_{ij} = 1) = 1 - \pi_{ij}$.

When f is a normal distribution, equation 6.7 can explicitly be written as follows.

$$\begin{gathered} Y_{ijl} = \beta_{0_l} + \beta_{Z_l} Z_j + \beta_{W_l} W_j + \beta_{X_l} X_{ij} + \upsilon_{0_{j_l}} + \epsilon_{ij_l} \\ C_{ijl} \sim \text{Bern}(\pi_{ijl}) \end{gathered} \tag{6.8}$$

where $\upsilon_{0_{j_l}} \sim N(0, \sigma^2_{\upsilon_{0l}})$, $\epsilon_{ij_l} \sim N(0, \sigma^2_{\epsilon_{0l}})$ and $\zeta_j \sim N(0, \sigma^2_\zeta)$. The LATE is given by β_{Z_2}. Under *exclusion restriction*, β_{Z_1} is constrained to be 0.

In the one-sided non-adherence setting, the true adherence class C_{ij} is unknown but it is assumed that the partially observed (*i.e.* observed in those assigned to treatment) binary variable R_{ij} is an indicator of C_{ij}. We denote $p_{ij_2} = P(R_{ij} = 1) = p_{ij}$ the estimated probability of being a *complier* and $p_{ij_1} = P(R_{ij} = 0) = 1 - p_{ij}$ the estimated probability of being a *never-taker*.

I present in the next section how to estimate the parameters of equation 6.8. This model can be fitted using either a Bayesian approach or a frequentist approach based on the EM algorithm.

6.3.3.1 Bayesian multilevel mixture model

The Bayesian multilevel mixture (BMM) model without baseline covariates is specified as follows.

$$\begin{cases} Y_{ijl} \sim N(\mu_{ij_l}, \sigma^2_{\upsilon_l} + \sigma^2_{\epsilon_l}) \ ; \ l \in \{1, 2\} \\ \mu_{ij_1} = \beta_{0_1} + \beta_{Z_1} Z_j + \upsilon_{j_1} \\ \mu_{ij_2} = \beta_{0_2} + \beta_{Z_2} Z_j + \upsilon_{j_2} \\ \upsilon_{jl} \sim N(0, \sigma^2_{\upsilon_l}) \\ R_{ij} \sim \text{Bern}(p_{ij}) \ ; \ \text{logit}(p_{ij}) = \lambda_0 + \zeta_j \ ; \ \zeta_j \sim N(0, \sigma^2_\zeta) \end{cases} \tag{6.9}$$

where $\sigma^2_{\upsilon_1}$ and $\sigma^2_{\upsilon_2}$ are the between-cluster variance in the *never-takers* and *compliers* classes, respectively and ζ_j is the random effect included in the logistic regression modelling the odds of being a complier. We assume, under exchangeability by design, that the estimated probability of being a complier in the active group is equal to

the true probability of being a complier *i.e.* $\pi_{ij} = p_{ij}$.

Adjustment for baseline covariates Covariates adjustment can easily be achieved by simply including baseline covariates W_j and X_{ij} in model 6.9, only if W_j and X_{ij} are fully observed. However, when there are some covariates with missing values, equation (6.10) shows how covariate adjustment is achieved. Let V_{1ij} and V_{2ij} be two baseline covariates with missing values such that V_{1ij} has less missing values than V_{2ij}. We assume here that the baseline covariates W_j and X_{ij} are fully observed. The BMM model with adjustment for baseline covariates (with and without missing values) is as follows.

$$\begin{cases} Y_{ijl} \sim N(\mu_{ij_l}, \sigma^2_{\upsilon_l} + \sigma^2_{\epsilon_l}) \ ; \ l \in \{1,2\} \\ \mu_{ij_l} = \beta_{0_l} + \beta_{Z_l} Z_j + \beta_{W_l} W_j + \beta_{X_l} X_{ij} + \upsilon_{j_l} \\ \upsilon_{j_l} \sim N(0, \sigma^2_{\upsilon_l}) \\ V_{1ij} \sim N(\theta_{1_{ij_l}}, \sigma^2_{\omega_{1_l}} + \sigma^2_{\xi_{1_l}}) \ ; \ V_{2ij} \sim N(\theta_{2_{ij_l}}, \sigma^2_{\omega_{2_l}} + \sigma^2_{\xi_{2_l}}) \\ \theta_{1_{ij_l}} = \gamma_{0,V_{1l}} + \omega_{1j_l} + \gamma_{Z,V_{1l}} Z_j + \gamma_{W,V_{1l}} W_j + \gamma_{X,V_{1l}} X_{ij} \\ \theta_{2_{ij_l}} = \gamma_{0,V_{2l}} + \omega_{2j_l} + \gamma_{Z,V_{2l}} Z_j + \gamma_{W,V_{2l}} W_j + \gamma_{X,V_{2l}} X_{ij} + \gamma_{X,V_{2l}} V_{1_{ij}} \\ \omega_{1j_l} \sim N(0, \sigma^2_{\omega_{1_l}}) \ ; \ \omega_{2j_l} \sim N(0, \sigma^2_{\omega_{2_l}}) \\ R_{ij} \sim \text{Bern}(\pi_{ij}) \ ; \ \text{logit}(\pi_{ij}) = \lambda_0 + \gamma_{\text{W}} W_j + \gamma_{\text{X}} X_{ij} + \zeta_j \ ; \ \zeta_j \sim N(0, \sigma^2_\zeta) \end{cases} \tag{6.10}$$

Under *exclusion restriction*, β_{Z_1} is set at 0, that is, the effect of treatment assignment on Y_{ij} among never-takers. A vague prior distribution (normally distributed and centered at 0) is often assumed for the coefficients. For level-2 variances in particular, Gelman [114] recommended starting with a non-informative uniform prior density on standard deviation parameters unless the number of clusters is low (below 5 for instance); otherwise, the uniform prior density tends to lead to high estimates of the standard deviation. He also suggested the use of a non-informative prior from the Half-Cauchy distribution on the scale of the standard deviation when several variance parameters are needed. The Half-Cauchy distribution is more flexible and behaves better for standard deviations near 0. Jeffrey's prior is often used for level-1 standard deviation.

I fit the model using Markov Chain Monte Carlo (MCMC) simulation [115, 116] in the "Just Another Gibbs Sampler" (JAGS) through R using the "Rjags" package [117]. I mainly assess the convergence of Markov chain by looking at the trace plots and formally using the Gelman-Rubin statistic, whose value is 1 for perfect mixing of chains [118]. It is recommended that a Gelman-Rubin statistic greater than 1.10 indicates poor mixing of chains [119].

6.3.3.2 Multilevel mixture model via expectation-maximization

The expectation-maximization (EM) algorithm is an iterative computation of maximum likelihood estimates that can be used to address incomplete-data problems [120]. By incomplete-data, it is meant that there exists some hidden variables or parameters not measured in the observed data. Here, the data are incomplete because the adherence class indicator R is missing for the control group.

It mainly consists of two steps, the expectation step (E-step) and the maximization step (M-step). In some settings like mixture modelling, the likelihood function of the incomplete-data is difficult to maximize while the likelihood of the complete-data is much simpler to work out [120]. To circumvent the maximization issue, it is assumed that the true adherence class C_{ij} for each individual unit and also the random effects υ_j are known. In the E-step, the posterior probabilities $\pi_{ij_l}^{(pos)}$ of π_{ij_l} are calculated after making an initial guess about the parameters $\boldsymbol{\beta}_l, \boldsymbol{\epsilon}_l^2,$ and $\boldsymbol{\upsilon}_{j_l}$. In the M-step, the expected subsequent likelihood is maximized after substituting $\pi_{ij_l}^{(pos)}$ with π_{ij_l}. The EM algorithm goes on iteratively until convergence. The multilevel mixture model via EM can be fitted in software such as Mplus [30, 109] or in R using the "lavaan" package [121] for instance.

Chapter 7.

Simulation study of individual-level LATE estimation in CRTs

7.1 Introduction

This chapter investigates, via simulations, the finite sample performance of TSLS, Wald and Bayesian multilevel mixture estimations of IL-LATE introduced in chapter 6. I consider a setting of one-sided non-adherence CRTs with moderate number of clusters (25 clusters per group) and a total size of $1,000$ units on average. The data generating process is the same as in chapter 5. Note that the simulations assume that the identification assumptions for LATE are met.

The current chapter is organized as follows. Section 7.2 summarises the analysis and criteria used to assess the methods' performance. Section 7.3 presents the simulation results on the performance of the Wald or conditional Wald, TSLS with HWR SEs, TSLS with Moulton-corrected SEs and Bayesian multilevel mixture modelling for estimating IL-LATE. Finally, section 7.4 summarizes the chapter. I do not include estimations via multilevel mixture EM because of inaccessibility to the specialized software Mplus [109].

The objectives of the simulations are to assess the performance of the methods mentioned above in terms of empirical bias and coverage, and to provide recommendations as to when and how analysts should implement those methods when interested in estimating IL-LATE.

7.2 Inclusion criteria, analysis and performance criteria

I implement the TSLS, Wald estimation and Bayesian multilevel mixture method introduced earlier in chapter 6. The analyses are performed on the 2500 simulated CRT data sets used in chapter 5, where the random treatment assignment is relevant (*i.e.* first stage $F-$statistic from TSLS ≥ 10). I restrict the performance assessment to the settings of moderate CRTs size ($J = 50$ *i.e.* 25 clusters per trial group) and where LATE size is large ($0.4SD$). For every simulation, estimation via the Bayesian method in particular is based on equation (6.9). A chain equal or longer than 50000 is run until convergence that is assessed using the Gelman-Rubin statistic which is suggestive of good convergence if lower than 1.10 [118, 119].

Standard errors for the frequentist-based methods, here TSLS and the Wald estimator, are obtained via HWR approach and Moulton's correction. I construct the 95% CIs for the Wald estimates using normal approximation bootstrapped-based CIs with 1500 replications, where both clusters and individuals within clusters are re-sampled. TSLS and Wald estimations are performed using Stata 15 and the Bayesian analyses implemented in JAGS through R using the "Rjags" package. The analysis codes are shown in appendix A.10. A summary of the analysis scenarios is given in Table 7.1.

Table 7.1: Overview of TSLS, Wald and Bayesian multilevel mixture estimations of IL-LATE in the simulation study

Methods	**Analyses features**	
TSLS	Covariate adjustment	None W_j X_{ij} W_j and X_{ij}
	SE estimation	HWR Moulton
Wald	Covariate adjustment	None W_j X_{ij} W_j and X_{ij}
	SE estimation	Bootstrap-based normal approximation[a]
Bayesian multi-level mixture	Covariate adjustment	None W_j X_{ij} W_j and X_{ij}
	SE estimation	Bayesian (2.5^{th}-97.5^{th} percentiles)

[a] Both clusters and individuals within clusters are re-sampled, with 1 500 replicates.
IL: individual level; HWR: Huber-White-Rogers; SE: standard error.

7.2.1 Estimation methods

I perform TSLS, the Wald estimator and the Bayesian multilevel mixture modelling introduced earlier in Chapter 6. For TSLS, SEs are estimated using either HWR method [70, 110] or Moulton's factor correction [76].

For the Wald estimation, the SE of the ratio is obtained via bootstrapping using 1500 replicates with both clusters and individuals within clusters re-sampled and 95% CIs constructed based on normal approximation. I used 1500 bootstrap replicates as most practitioners suggested a number of replicates between 1000 and 2000.

As regards to the Bayesian multilevel mixture modelling, I fit equations shown in section 6.3.3.1. Uninformative normal priors are used for all coefficients and a vague inverse gamma prior for level-1 variance. I used Uniform and half-Cauchy priors for the level-2 standard deviation as recommended by Gelman [114].

7.2.2 Performance criteria

The performance criteria used are the same in chapter 5, that is, the empirical bias and coverage rates of the 95% CIs over 2500 replicate data sets per scenario. The bias uncertainty is presented using a 95% CI constructed based on the Monte Carlo Error. The acceptable coverage rate due to the finite sampling error for this number of replications is between 94.1% and 95.9%. For the Bayesian multilevel mixture, the median value is used as the point estimate of IL-LATE. Details on the criteria performance are presented in section 5.3.2.

7.3 Results

This section summarizes the performance of the Wald (conditional) estimator with bootstrapped SEs, TSLS with HWR or Moulton-corrected SEs and Bayesian multilevel mixture modelling with Uniform or half-Cauchy prior for the level-2 standard deviation when estimating IL-LATE. I present below the results for CRTs where adherence is at the cluster level and then when adherence is at the individual level.

7.3.1 Adherence at cluster level

Figure 7.1 displays the results of the estimation of IL-LATE reported in terms of empirical bias and coverage of the 95% nominal CI for the Wald estimator, TSLS estimation with HWR or Moulton-corrected SEs and the Bayesian multilevel mixture modelling with Uniform or Cauchy prior for the level-2 standard deviation when adherence is at cluster level. Panels A and B present the performance of these methods when the ICC for the outcome is 0.05 (low ICC) and 0.20 (high ICC), respectively.

The Wald, TSLS with HWR SEs and TSLS with Moulton SEs provide unbiased IL-LATE estimate with good coverage and outperform the Bayesian multilevel mixture modelling whether the ICC for outcome is low or high. TSLS with HWR SEs and TSLS with Moulton SEs have very similar performance, regardless of how covariate adjustment is done. The Wald (or conditional Wald) estimation with bootstrapped SEs is conservative *i.e.* has coverage above the 95% nominal CI when the ICC for the outcome is low. For the higher ICC, the Wald estimation like TSLS with HWR SEs and TSLS with Moulton-corrected SEs has a coverage at the 95% nominal level.

The Bayesian multilevel mixture modelling provides downward biased IL-LATE estimates, whether Uniform or half-Cauchy prior is used for the level-2 standard deviation. Regardless of the level of ICC for the outcome (as low as 0.05% or as high as 0.20%), the empirical bias is attenuated (about 3% to 5%) when no covariate adjustment is done whereas the bias increases (up to $15\% - 20\%$) with cluster-level and/or individual-level covariate adjustment. The increase in parameters to be estimated for the covariates while the number of clusters for predicting the adherence classes is relatively small (as only 25 clusters in the active group provides information) may reduce the model's predictive accuracy of adherence classes, which in turn may explain the bias of IL-LATE estimates after covariate adjustment. In the presence of low ICC for the outcome, the coverage is above the 95% nominal level when no covariate adjustment is done but falls below the 95% nominal level (between 89% and 93%) with cluster-level and/or individual-level covariates adjustment. How-

ever, in the presence of high ICC for the outcome, the Bayesian multilevel mixture is conservative with and without covariate adjustment. Having only 25 clusters in the active group seems too low to model the adherence classes. Thus, the Bayesian multilevel mixture model may require larger number of clusters in the active group to perform well.

7.3.2 Adherence at individual level

Figure 7.2 presents the results of the estimation of IL-LATE reported in terms of the empirical bias and coverage of the 95% CI when the Wald (or conditional Wald) estimator, TSLS with HWR SEs, TSLS with Moulton-corrected SEs and Bayesian multilevel mixture modelling with uninformative Uniform or Cauchy prior for the level-2 standard deviation are used when adherence is at the individual level. Panels A and B show the performance of these estimation methods when the ICC for outcome is 0.05 (low ICC) and 0.20 (high ICC), respectively.

When the ICC for outcome is low, the Wald estimator and TSLS with HWR or Moulton's SEs lead to unbiased IL-LATE estimates. This applies to the Bayesian multilevel mixture modelling except when a half-Cauchy prior is used for the level-2 standard deviation and only the individual-level covariate is adjusted for. The Wald estimator with bootstrapped SEs has good coverage except when only individual-level covariate is adjusted for while there is a cluster-level covariate strongly affecting both outcome and adherence to treatment. In the latter, the coverage is slightly below the 95% nominal level ($\approx$ 93.5%). TSLS with Moulton's SEs correction shows good coverage irrespective of how covariate adjustment is done. However, TSLS with HWR SEs has a coverage slightly below the 95% nominal level ($\approx$ 93.5%). The coverage for the Bayesian multilevel mixture modelling is good but tends to be conservative especially when a cluster-level covariate strongly associated with outcome and adherence is adjusted for. The Bayesian multilevel mixture modelling has a coverage close to the 95% nominal and is less conservative when no covariate adjustment is done or when the cluster-level covariate has a small effect on outcome and adherence to treatment.

In the presence of higher ICC for the outcome, the Wald estimator and TSLS with HWR or Moutlon's SEs lead to unbiased IL-LATE unlike the Bayesian multilevel mixture modelling that shows little downward bias (around 1%). The bias is attenuated when both the cluster-level and individual-level covariates have strong effects on outcome and adherence to treatment, and are adjusted for. The coverage for Wald with bootstrapped SEs, TSLS with HWR SEs and TSLS with Moulton-corrected SEs is below the 95% nominal (between 92.5% and 94%) and gets close to the 95% nominal when the cluster-level covariate only has little effect on both outcome and adherence. The Bayesian multilevel mixture model, however, shows good coverage at the 95% nominal level.

7.4 Summary

The current chapter compares the performance of the Wald (or conditional Wald) estimator, TSLS with HWR SEs, TSLS with Moulton SEs and the Bayesian multilevel mixture modelling in terms of empirical bias and coverage at the 95% nominal CI level when estimating IL-LATE under the required identification assumptions in CRTs with moderate number of clusters (here 25 clusters per trial group) and where there is one-sided non-adherence at the cluster level or at the individual level.

When adherence is at the cluster level, the Wald, TSLS with HWR SEs and TSLS with Moulton SEs provide unbiased IL-LATE estimate with good coverage but the Wald estimation is slightly conservative when the ICC for outcome is low. TSLS with HWR SEs and TSLS with Moulton SEs show very similar performance, whether covariates are adjusted for or not. The Wald, TSLS with HWR SEs and TSLS with Moulton SEs outperform the Bayesian multilevel mixture modelling irrespective of the level of ICC for outcome (low or high).

However, in the presence of individual-level adherence, the Bayesian multilevel mixture modelling outperforms the Wald estimator and TSLS with HWR or Moulton's SEs in terms of coverage. TSLS with Moulton-corrected SEs performs well in terms of empirical bias and coverage when the ICC for the outcome is low. Unlike the Wald and TSLS, IL-LATE estimates from the Bayesian multilevel model show little

but negligible bias (below 1%).

Irrespective of the level of ICC for the outcome or strength of association between cluster-level or individual-level covariate and the outcome or adherence to treatment, avoiding covariate adjustment appears as a good, simple and safer option when estimating IL-LATE in the settings considered in these simulations *i.e.* in settings where covariates are not used as design factors (in stratified randomisation, for example). However, although not covered in our simulations, it is sensible to adjust at least for covariates used as design factors. For the Bayesian multilevel mixture model, the Uniform prior for the standard deviation appears to perform well.

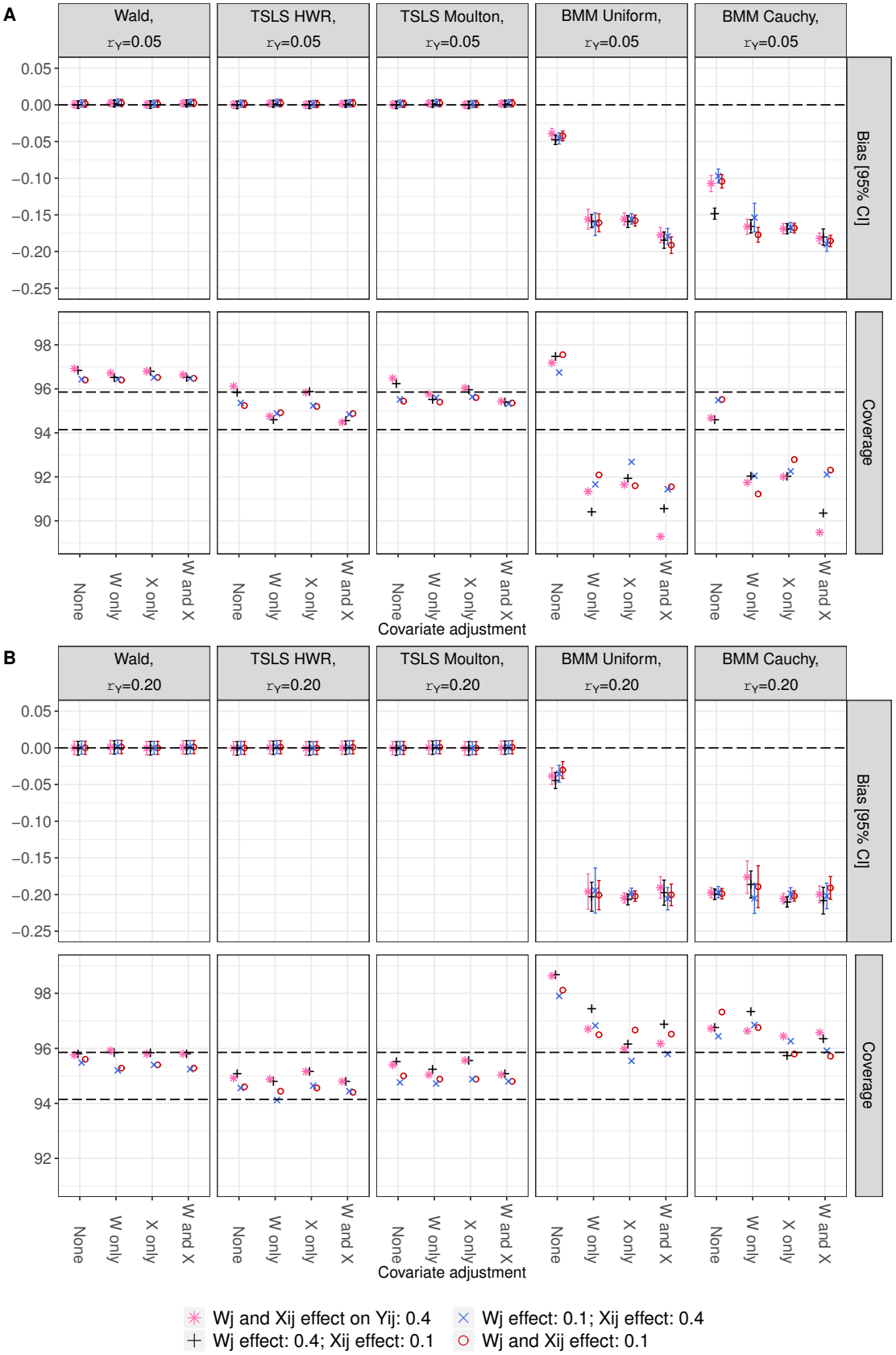


Figure 7.1: Performance of Wald, TSLS and Bayesian multilevel mixture methods to estimate individual-level LATE in the presence of one-sided non-adherence at cluster level for CRT with 25 clusters per group where ICC for outcome is 0.05 (A) and 0.20 (B). The true LATE is 0.4 standard deviation. The long-dashed black parallel lines in the last panel are the acceptable 95% coverage range.

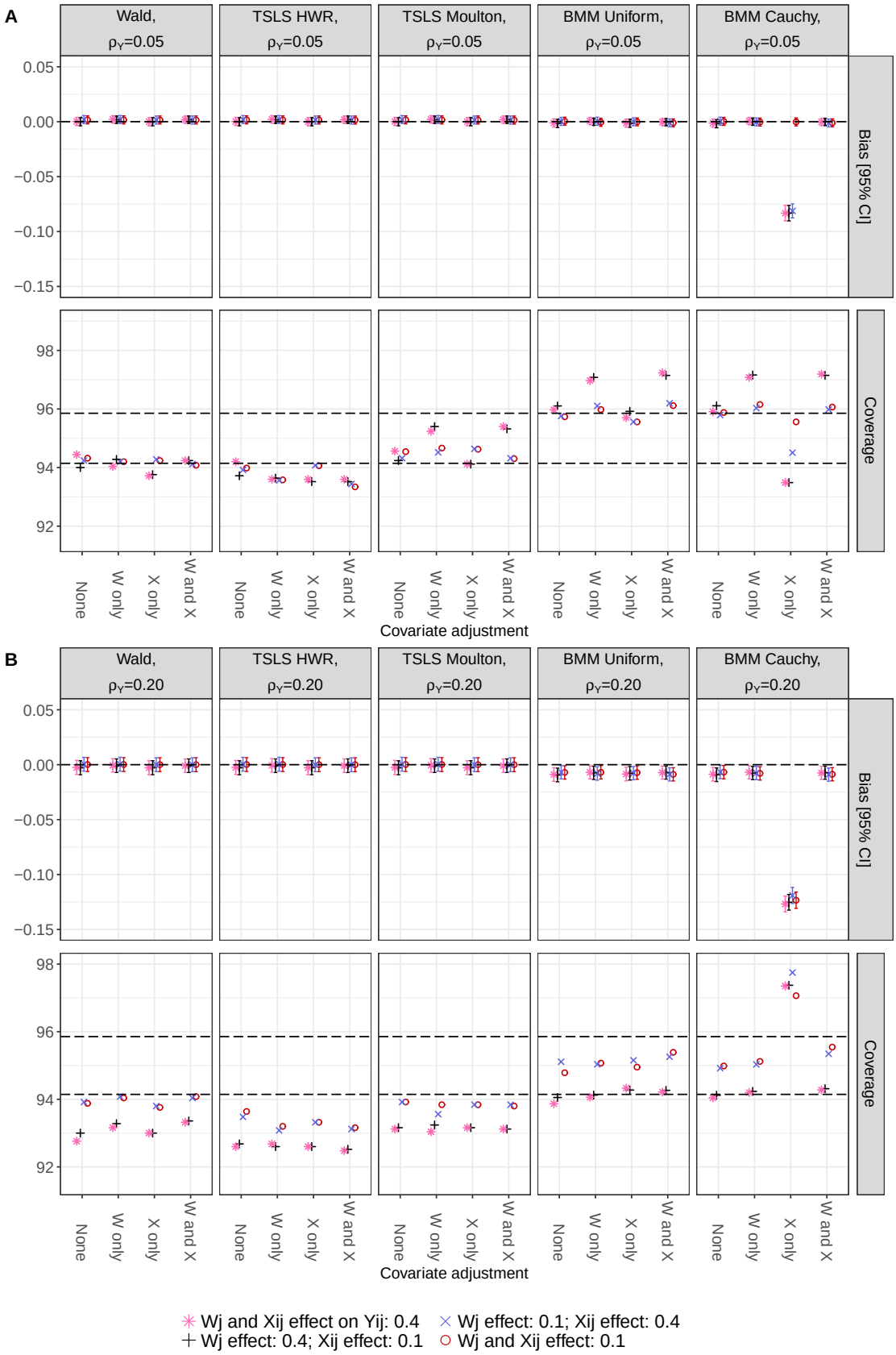


Figure 7.2: Performance of Wald, TSLS and Bayesian multilevel mixture methods to estimate individual-level LATE in the presence of one-sided non-adherence at individual level for CRT with 25 clusters per group where ICC for outcome is 0.05 (A) and 0.20 (B). The true LATE is 0.4 standard deviation. The long-dashed black parallel lines in the last panel are the acceptable 95% coverage range.

Chapter 8.

Illustration of LATE estimation at the cluster and individual level using the OPERA and TXT4FLUJAB trial data

8.1 Introduction

The chapter presents the re-analysis of the TXT4FLUJAB and OPERA trials introduced earlier in sections 1.5.1 and 1.5.2, respectively. Both cluster-level LATE (chapter 4) and individual-level LATE (chapter 6) are the estimands of interest.

The TXT4FLUJAB trial investigators were interested in the causal treatment effects at the cluster level (here GPs) expressed as a mean risk difference, where the outcome is a binary variable indicating whether patients received the influenza vaccination or not. Adherence to treatment was binary, two-sided and occurred at the individual level (here patients). Patients whose GPs were allocated to the active group are said to adhere if they received the reminder text message; whereas, those whose GPs were allocated to the control group are said to adhere if they did not receive any reminder text messages. Missing data were not an issue as everything was automatically recorded.

Regarding the OPERA trial, investigators intended to assess causal treatment effects, but the trial protocol did not formally mention how this would be done. The investigators anticipated and collected information on whether residents attended any exercise sessions. There was no formal definition for adherence to treatment. For illustrative purposes, I adopt a working definition of all-or-none adherence. Residents in the active group are considered to have received their intended treatment if they attended at least one group exercise session, whereas those in the control

group received their intended treatment if they did not attend any group exercise sessions. As little exposure (at least one) to the exercise sessions may improve resident's outcome than no exposure at all, the all-or-none definition of adherence is less susceptible to violate the ER assumption than considering some residents with a non-zero threshold of number of exercise sessions to not benefit at all from their exposure to the active treatment. The control group did not have access to the exercise sessions. The outcome of interest is the "short physical performance battery score (SPPB)" at 12 months, treated as a continuous variable as in the original publication [37].

As missing data were present in the OPERA trial, I used multilevel joint modelling multiple imputation [122, 123] to handle missing data assuming missing at random, that is, the distributions of the missing and observed data are the same, conditional on covariates. The multilevel multiple imputation allows for clustering of the individual records according to the residential home and is carried out using the "jomo" package in R [124], separately in the control and active groups. The multilevel imputation model includes the actual treatment received as a covariate, enabling us to allow non-adherence to predict missing values. The "jomo" package offers the advantage of – (i) imputing with a common level-1 covariance matrix across level-2 units, – (ii) imputing with a cluster-specific level-1 covariance matrices, and – (iii) imputing while allowing for the level-1 covariance matrix to be randomly distributed across level-2 units [123–125].

The present chapter is organized as follows. Section 8.2 presents a descriptive analysis, the cluster-level and individual-level ITT and LATE estimations for the OPERA trial. Section 8.3 replicates similar analyses for the TXT4FLUJAB trial. Section 8.4 summarises the findings. The analyses presented here assume that all IV assumptions and monotonicity at the individual and cluster levels are met. Sensitivity analyses to assess departures from these assumptions are presented in chapter 9.

8.2 Re-analysis of the OPERA trial

I present here the descriptive analysis and the ITT and LATE estimates at the cluster and individual levels.

8.2.1 Descriptive analysis

The OPERA trial was a non-blinded CRT carried out in 78 care-homes, of which 35 were allocated to the active group and 43 to the control group. Recall that the intervention in the active group was a complex programme involving training on depression awareness for care home staff, 45 minutes physiotherapist-led group exercise sessions for residents (delivered twice a week) and a whole home component designed to motivate residents to increased daily physical activity. Care homes staff in the control group only received the training on depression awareness [37]. Care homes represent the clusters and residents are the individuals.

The marginal ICC of the outcome was 0.09 overall, 0.02 in the control group and 0.11 in the active group, suggestive of a potential presence of level-1 and/or level-2 variance heterogeneity across trial groups. Table 8.1 shows the care-homes' characteristics and residents' characteristics at baseline as well as the percentages of adherence to treatment in each trial group. In total, 900 residents were enrolled (498 in the control group and 402 in the active group). The distribution of the clusters size (number of residents within clusters) was similar across trial groups; the median (range) clusters size was 6 (2-11) and 6 (2-15) in the control and active group, respectively. However, the distribution of cluster sizes was left-skewed with a mean cluster size of 12 in each group, pointing out some imbalance in the cluster sizes.

Care homes and residents characteristics were balanced across trial groups at baseline. The percentages of missing values were similar across groups and less than 1% for baseline antidepressant uptake and age. However, the percentage of missing values was higher in the control group compared to the active group for "moderate to severe cognitive impairment score" (MMSE) (19% vs. 14%) and SPPB (17% vs. 13%) at baseline.

Based on our working definition of adherence to treatment at individual level, all residents (100%) assigned to the control group received the intended treatment whereas 89% of residents allocated to the active group received the intended treatment. The median percentage (range) of active treatment uptake at the care home level was 92% (57%-100%) in the active group; the percentage of active treatment uptake at the care home level was 0% for all care-homes in the control group.

Table 8.1: Baseline characteristics and percentages of treatment received by trial group

Characteristics	**Control**	**Active**
Number of care-homes, n (%)	43	35
Total number of residents	498	402
Number of resident per care home, median (range)	6 (2-11)	6 (2-15)
Location		
London, n (%)	204 (41)	208 (52)
Warwick, n (%)	294 (59)	194 (48)
Type of care home provider		
Private & care home, n (%)	280 (56)	172 (43)
Private & Nursing, n (%)	117 (23)	125 (31)
Voluntary, n (%)	101 (20)	105 (26)
Size of care home		
$<$ 32 residents, n (%)	250 (50)	204 (51)
$\geq$ 32 residents, n (%)	248 (50)	198 (49)
Gender		
Female, n (%)	388 (78)	298 (74)
Male, n (%)	110 (22)	104 (26)
Antidepressant treatment		
Yes, n (%)	158 (31.7)	111 (27.6)
No, n (%)	337 (67.7)	290 (72.1)
Missing, n (%)	3 (0.6)	1 (0.3)
Residents' age in years at baseline		
median (range)	88 (65-107)	87 (65-107)
Missing, n (%)	4 (0.8)	3 (0.7)
Residents' MMSE at baseline		
median (range)	18 (0-30)	19 (0-30)
Missing, n (%)	94 (19)	56 (14)
Residents' SPPB at baseline		
median (range)	1 (0-10)	1 (0-10)
Missing, n (%)	84 (17)	54 (13)
Received intended treatment, n (%)	498 (100)	356 (89)

Continued on next page

Table 1 Continued

Characteristics	Control	Active
Percentage of active treatment uptake at the care home level, median (range)	0 (0-0)	92 (57-100)

The variables dictionary is as follows: – the outcome is SPPB at 12 months since enrolment (*SPPB2*), continuous and measured at the individual level, – the individual-level covariates are baseline SPPB (*SPPB0*, continuous), baseline "moderate to severe cognitive impairment score" (*MMSE0*, continuous), age at baseline (*AGE*, continuous), sex (*SEX*, binary) and baseline antidepressive treatment (*ANTIDEP0*, binary), – the cluster-level covariates: care home location (*PLACE*, binary), size of care home (*SIZE*, binary) and type of home (*HOME*, categorical with three levels), and – the random treatment allocation of care-homes (*ALLOC*, binary). As often done in practice, those covariates were pre-selected by the trial investigators at the design stage as variables to be included in the model for estimating the treatment effects. I therefore adjust the treatment effects for those covariates.

8.2.2 Cluster-level summary analyses

The covariates to be adjusted for are (i) individual-level covariates: *AGE*, *SEX*, *SPPB0* and *ANTIDEP0*, (ii) cluster-level covariates: *PLACE*, *SIZE* and *HOME* and the proportions of residents with MMSE at baseline (*MMSE0*) < 20 within care-homes [37]. The proportions of residents with *MMSE0* < 20 within care-homes as per [37] were computed ignoring the presence of missing values in *MMSE0*. This may misrepresent the true proportions of residents with MMSE < 20 within care-homes and subsequently introduced measurement error issues, unless the missingness is completely at random or the missing values at the residents level are adequately handled. For simplicity, I use the individual-level baseline MMSE instead of the proportions of residents with MMSE < 20.

I re-analyse the OPERA trial using cluster-level (CL) summary approaches, with and without covariate adjustment. I present both complete records analysis and analysis from the multilevel multiple imputation. I gradually increased the number

of imputations in order to choose an adequate number of simulations that leads to a Monte Carlo error (MCE) below 1% for the point estimates, SEs and lower/upper limits of 95% CI. I finally generated 250 imputed datasets and the estimates are pooled using the Rubin's rule [126,127]. Variables included in the imputation model are those listed earlier. There was no variable used as an auxiliary in the imputation model. Thus, the complete record and multiple imputation analysis model include the same variables.

I applied the CL-summary approaches introduced in chapter 3, using the unadjusted and adjusted CL-means *SPPB2* as outcome, without/with weighting and without/with covariate adjustment. Both complete records analysis and multilevel multiple imputation are reported. These analyses are performed using Stata 15, whereas data are imputed using the "jomo" package in R [124]. The multilevel multiple imputation and the analysis codes are in Appendix A.11.

Prior to individual-level ITT analyses, I conducted an exploratory analysis to investigate the residual variance heterogeneity and understand to what extent covariate adjustment and Huber-White (HW) SEs may affect the ITT inferences.

Exploratory assessment of residual variance

Chapter 3 introduced how the unadjusted and adjusted CL summaries are computed. The unadjusted CL summaries here are the CL-means *SPPB2* whereas the adjusted CL summaries are the CL-means of the residuals from OLS regression of *SPPB2* on individual-level covariates *SPPB0*, *MMSE0*, *SEX* and *ANTIDEP0*. For individual-level analyses, *SPPB2* is analysed using a mixed effects linear regression. Table 8.2 shows the residual outcome variance for each trial group and the residual outcome variance from the ITT analyses, with and without covariates using complete records without weighting.

As expected, covariate adjustment reduces the residual variance. There is a noticeable variance heterogeneity across trial groups. This is attenuated by the inclusion of covariates associated with the outcome. For the adjusted CL-summary analyses, there is no apparent suggestion of residual variance heterogeneity, likely due to

the lack of evidence of association between the outcome and cluster-level covariates (*PLACE*, p=0.25; *SIZE*, p=0.28; *HOME*, p=0.81). It is worth noting that the use of adjusted CL-summary outcome greatly reduces the residual variance as opposed to the unadjusted CL-summary analysis. This reduction of residual variance obtained from adjusted CL-summary analysis would provide some efficiency gain.

Table 8.2: Residuals variance of SPPB at the individual level, unadjusted and adjusted CL-summaries (means) SPPB at 12 months by trial group and overall, using complete records analyses without weighting

	Unadjusted analysis			**Adjusted analysis[a]**		
	ITT[b]	**Control**	**Active**	**ITT[b]**	**Control**	**Active**
Individual-level *SPPB2*						
Level-1 variance	4.153	3.512	4.931	2.090[c]	1.986[c]	2.225[c]
Level-2 variance	0.384	0.091	0.671	0.226[c]	0.213[c]	0.263[c]
Unadjusted CL-means *SPPB2*						
Residuals variance	1.160	0.932	1.441	0.970	0.961[e]	0.845
Adjusted CL-means *SPPB2*[d]						
Residuals variance	0.631	0.640	0.619	0.631[e]	0.667[e]	0.614[e]

[a] Adjusted for cluster-level covariates *PLACE*, *SIZE* and *HOME*.

[b] Pooled residual variance obtained from the intention-to-treat (ITT) analysis.

[c] Also adjusted for individual-level covariates *SPPB0*, *MMSE0*, *AGE*, *SEX* and *ANTIDEP0*.

[d] Residuals obtained from OLS regression with individual-level covariates *SPPB0*, *MMSE0*, *AGE*, *SEX* and *ANTIDEP0*.

[e] No evidence of association between outcome and any of the cluster-level covariates.

CL-ITT estimates

The cluster size imbalance previously observed may lead to varying precision of the CL-means *SPPB2*, and may potentially cause heteroscedasticity. I used unweighted and weighted linear regressions to estimate CL-ITT effect, expressed as the mean difference in CL-means *SPPB2*. The weights are cluster size (CS) and minimum variance (MV). HW SEs and/or the weighting are used to circumvent the suspected heteroscedasticity. No small degrees of freedom adjustment is needed because OLS inference is based on t-distribution.

Tables 8.3 and 8.4 show the CL-ITT effect estimates when the unadjusted and ad-

justed CL-means *SPPB2* are analysed, respectively. The CL-ITT effect estimates vary considerably across analysis methods and increase after CL covariates adjustment. Nevertheless, the conclusions from these results point towards no evidence of ITT effect at the cluster level, except when adjusted CL-means *SPPB2* are analysed without weighting and using complete records where there was weak evidence of ITT effect at the cluster-level (unadjusted ITT effect: 0.30, 95% CI: -0.06,0.66, p=0.098 and adjusted ITT effect: 0.32, 95% CI: -0.06,0.70, p=0.094 when HW SEs are used). This efficiency gain from analysing the adjusted CL-means *SPPB2* with no weighting is due to the substantial reduction of the residual variance highlighted in section 8.2.2, as opposed to the unadjusted CL-means *SPPB2*.

For all analysis methods, the CL-ITT point estimates using data from multilevel multiple imputation are lower than those obtained from complete records analysis. The multiple imputation leads to more conservative inferences, suggesting no evidence of CL-ITT effect. As expected, there is some efficiency gain when covariates are adjusted for, regardless the methods. This efficiency gain is, in fact, due to the reduction in the residual variance induced by including a CL covariate that is associated with the CL-means *SPPB2* (*HOME*, p<0.001), and including individual-level covariates associated with *SPPB2* at the individual level (*SPPB0*, p<0.001; *MMSE0*, p=0.04 and *SEX*, p=0.03).

The HW SEs have little impact when using adjusted CL-summary outcome and not including any CL covariates in the regression model, regardless of the weighting strategy and whether data are imputed or not. There was a similarity between the residual variance across trial groups when analysing the adjusted CL-means *SPPB2* without weighing (section 8.2.2). This is possibly but not necessarily an indication of homoscedastic residuals, which is supported by the "absolute" no change of the 95% CIs when using HW SEs with no weighting (Table 8.4, 95% CI: -0.057 to 0.663 assuming homoscedasticity and 95% CI: -0.057 to 0.662 assuming heteroscedasticity). The use of *CS* or *MV* weights influence the CL-ITT effect estimates. *MV* weighting may be preferable as providing the optimal variance in principle. In addition, from our results, ITT effect estimates when using *MV* weights often fall

within the estimates from not using any weights and *CS* weighting, offering then a reasonable compromise.

Table 8.3: Care home-level ITT effect estimates (as mean difference) on SPPB at 12 months, using unadjusted CL-summaries on complete records and multiple imputed data

		Unadjusted		**Adjusted[a]**	
		ITT (95% CI)	**p**	**ITT (95% CI)**	**p**
Complete records					
No weighting	None	0.292 (-0.196, 0.780)	0.237	0.369 (-0.082, 0.821)	0.107
	HW	(-0.207, 0.791)	0.247	(-0.088, 0.826)	0.112
Cluster size	None	0.327 (-0.131, 0.785)	0.160	0.370 (-0.067, 0.806)	0.096
weights	HW	(-0.144, 0.797)	0.171	(-0.085, 0.825)	0.110
Minimum-variance	None	0.305 (-0.170, 0.779)	0.205	0.366 (-0.079, 0.811)	0.106
weights	HW	(-0.176, 0.785)	0.211	(-0.086, 0.818)	0.111
Multilevel multiple imputation					
No weighting	None	0.286 (-0.184, 0.755)	0.233	0.355 (-0.082, 0.792)	0.112
	HW	(-0.200, 0.771)	0.249	(-0.096, 0.806)	0.123
Cluster size	None	0.261 (-0.185, 0.708)	0.160	0.309 (-0.119, 0.736)	0.157
weights	HW	(-0.198, 0.721)	0.171	(-0.133, 0.750)	0.171
Minimum-variance	None	0.273 (-0.187, 0.733)	0.244	0.333 (-0.100, 0.766)	0.131
weights	HW	(-0.198, 0.744)	0.256	(-0.110, 0.777)	0.141

HW: Huber-White. [a] Adjusted cluster-level covariates: *PLACE*, *SIZE* and *HOME*.

Table 8.4: Care home-level ITT effect estimates (as mean difference) on SPPB at 12 months, using adjusted CL-summaries on complete records and multiple imputed data

		Unadjusted		**Adjusted[a]**	
		ITT (95% CI)	**p**	**ITT (95% CI)**	**p**
Complete records					
No weighting	None	0.303 (-0.057, 0.663)	0.098	0.320 (-0.044, 0.684)	0.084
	HW	(-0.057, 0.662)	0.098	(-0.056, 0.696)	0.094
Cluster size	None	0.238 (-0.098, 0.573)	0.162	0.275 (-0.067, 0.616)	0.114
weights	HW	(-0.093, 0.568)	0.156	(-0.071, 0.620)	0.118
Minimum-variance	None	0.266 (-0.083, 0.615)	0.133	0.292 (-0.063, 0.646)	0.105
weights	HW	(-0.075, 0.607)	0.125	(-0.067, 0.650)	0.109
Multilevel multiple imputation					
No weighting	None	0.236 (-0.127, 0.599)	0.203	0.251 (-0.114, 0.617)	0.178
	HW	(-0.130, 0.601)	0.206	(-0.122, 0.624)	0.187
Cluster size	None	0.181 (-0.166, 0.527)	0.307	0.217 (-0.134, 0.568)	0.226
weights	HW	(-0.168, 0.530)	0.310	(-0.140, 0.573)	0.233
Minimum-variance	None	0.209 (-0.147, 0.566)	0.250	0.232 (-0.127, 0.592)	0.205

Continued on next page

Table Continued

		Unadjusted ITT (95% CI)	p	Adjusted[a] ITT (95% CI)	p
weights	HW	(-0.146, 0.564)	0.248	(-0.130, 0.595)	0.209

HW: Huber-White. [a] Adjusted cluster-level covariates: *PLACE*, *SIZE* and *HOME*.

8.2.3 Cluster-level LATE

In this section, I estimate LATE of attending at least one group exercise session on *SPBB2* at the cluster and individual levels. Analyses are performed on complete records and after multilevel multiple imputation. Before LATE estimation, I discuss the plausibility of LATE identification assumptions.

8.2.3.1 Plausibility of LATE identification assumptions

By design, the randomised treatment is unconfounded. The relevance of randomisation as instrument (in the first stage) is assessed based on the rule of thumb by Staiger and Stock [106] ($F(1, 76) = 2505.17 > 10$), suggesting that treatment assignment is a relevant instrument.

Assuming exclusion restriction at the individual level implies that there is no other path through which a resident's *SPBB2* can be affected except via the exercise sessions. This assumption is disputable as residents whose care-homes are assigned to the control group may feel discriminated and get demotivated during the assessment of their physical function. Thus, despite not actually attending any exercise sessions, such demotivated residents may have their *SPBB2* affected by the home care random treatment allocation.

The monotonicity assumption (that there are no defiers) is met as care-homes assigned to the control group did not implement any exercise sessions (see Table 8.9). Although care-homes are recruited in such a way to minimise interference across homes, it is not excluded that residents from different care-homes interact and influence each others' potential outcome.

8.2.3.2 CL-LATE estimates

As introduced in chapter 4, the Wald estimator with the Schochet-Chiang SEs (referred to as Schochet-Chiang method or estimator) is used and TSLS without and with weighting (*CS* or *MV* weights) are fitted using unadjusted or adjusted CL summary outcome. The treatment received is the unadjusted CL-proportions of residents who received the treatment *i.e.* attended at least one group exercise session within the cluster.

I include the same baseline covariates as in the ITT analyses and assume homoscedasticity or heteroscedasticity for TSLS and Schochet-Chiang estimation. For Schochet-Chiang method, I assume homoscedasticity and conduct analyses on complete records only. For TSLS, in addition, small sample degrees of freedom (SSDF) adjustment and normal approximation are considered. Analyses are performed in Stata 15 and the codes can be found in Appendix A.11.

Table 8.5 reports the unadjusted CL summary analysis results, while Table 8.6 reports the adjusted CL summary analysis results. Covariates adjustment provide some efficiency gain regardless the methods, whether complete case analysis or multiple imputation is used. As for CL-ITT effect estimates, we note that adjusting for covariates increases the CL-LATE point estimates. CL-LATE estimates from the multilevel multiple imputation are more conservative.

As pointed out in chapter 4, the Schochet-Chiang Wald-based estimates are equivalent to the unweighted CL-TSLS estimates when there is no covariate adjustment. This translates in the results not only without covariate adjustment but also when CL covariates are adjusted for. The 95% CIs slightly vary from one method to another but the Shochet-Chiang method and unweighted TSLS result in similar conclusions of no evidence of group exercise session effect on *SPPB2*, whether adjusted for CL covariates or not.

However, TSLS with *CS* and *MV* weighting in complete records analyses with CL covariates adjustment are suggestive of weak evidence of a positive group exercise sessions effect on *SPPB2*, except for TSLS with *CS* weights without any correction

for the SEs where the results suggest some evidence of a positive group exercise sessions effect on *SPPB2*.

Unsurprisingly, SSDF adjustment has in general a greater influence on the SEs than does the HW method. This influence is more pronounced when CL covariates are adjusted for, and is probably because of the loss of four degrees of freedom induced by the two binary covariates and one covariate with three levels in the presence of a moderate number of clusters ($J = 78$).

Considering the suspected presence of heteroscedasticity and the likely well estimated between-cluster variance because of the moderate number of clusters, CL-LATE obtained from TSLS with MV weights, SSDF adjustment and HW SEs when the outcome is the adjusted CL-means *SPPB2*, may seem appropriate. Thus, the CL-LATE estimate was 0.31 (95% CI: -0.08, 0.70; p=0.11) in complete records analyses and 0.26 (95% CI: -0.14, 0.67; p=0.21) for two-level multiple imputation analyses, both suggesting no evidence of causal treatment effect on *SPPB2* among care-homes whose residents were more likely to attend at least one group exercise session if offered (p=0.11).

Table 8.5: CL-LATE estimates (as mean difference) at care home level, of residents' attendance to at least one group exercise session on SPPB at 12 months, using unadjusted CL-summaries on complete records and multiple imputed data

		Unadjusted		**Adjusted[a]**	
		LATE (95% CI)	**p**	**LATE (95% CI)**	**p**
Complete records					
Schochet	None	0.314 (-0.218, 0.846)	0.248	0.394 (-0.106, 0.894)	0.122
TSLS					
No weighting	None	0.314 (-0.212, 0.839)	0.242	0.394 (-0.086, 0.874)	0.108
	HW	(-0.223, 0.851)	0.252	(-0.096, 0.884)	0.115
	SSDF	(-0.227, 0.855)	0.251	(-0.114, 0.903)	0.127
	SSDF + HW	(-0.239, 0.866)	0.262	(-0.125, 0.913)	0.134
CS weights	None	0.401 (-0.089, 0.890)	0.108	0.470 (0.010, 0.929)	0.045
	HW	(-0.101, 0.902)	0.117	(-0.014, 0.953)	0.057
	SSDF	(-0.103, 0.904)	0.117	(-0.017, 0.956)	0.058
	SSDF + HW	(-0.116, 0.917)	0.126	(-0.043, 0.982)	0.072
MV weights	None	0.368 (-0.134, 0.870)	0.150	0.447 (-0.019, 0.913)	0.060
	HW	(-0.143, 0.879)	0.158	(-0.031, 0.925)	0.067
	SSDF	(-0.148, 0.884)	0.160	(-0.047, 0.940)	0.075

Continued on next page

Table Continued

		Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p
	SSDF + HW	(-0.158, 0.894)	0.168	(-0.059, 0.953)	0.083
Multilevel multiple imputation					
Schochet	None	0.320 (-0.201, 0.841)	0.229	0.396 (-0.088, 0.881)	0.108
TSLS					
No weighting	None	0.320 (-0.197, 0.836)	0.225	0.396 (-0.073, 0.866)	0.098
	HW	(-0.214, 0.854)	0.241	(-0.086, 0.879)	0.108
	SSDF	(-0.202, 0.841)	0.230	(-0.202, 0.841)	0.230
	SSDF + HW	(-0.220, 0.860)	0.246	(-0.102, 0.895)	0.119
CS weights	None	0.295 (-0.201, 0.791)	0.295	0.347 (-0.115, 0.808)	0.141
	HW	(-0.214, 0.804)	0.256	(-0.129, 0.823)	0.154
	SSDF	(-0.206, 0.797)	0.249	(-0.130, 0.823)	0.154
	SSDF + HW	(-0.220, 0.810)	0.261	(-0.145, 0.838)	0.167
MV weights	None	0.307 (-0.201, 0.815)	0.236	0.373 (-0.093, 0.840)	0.117
	HW	(-0.213, 0.826)	0.247	(-0.103, 0.850)	0.125
	SSDF	(-0.206, 0.820)	0.241	(-0.108, 0.855)	0.129
	SSDF + HW	(-0.218, 0.832)	0.252	(-0.119, 0.866)	0.137

[a] Adjusted for place, care home size and care home type.

HW: Huber-White; SSDF: small sample degrees of freedom; CS: cluster size; MV: minimum-variance.

Table 8.6: CL-LATE estimates (as mean difference) at care home level, of residents' attendance to at least one group exercise session on SPPB at 12 months, using adjusted CL-summaries on complete records and multiple imputed data

		Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p
Complete records					
TSLS					
No weighting	None	0.330 (-0.051, 0.711)	0.090	0.348 (-0.026, 0.723)	0.068
	HW	(-0.050, 0.710)	0.089	(-0.037, 0.734)	0.077
	SSDF	(-0.062, 0.722)	0.098	(-0.048, 0.745)	0.084
	SSDF + HW	(-0.061, 0.721)	0.097	(-0.060, 0.775)	0.093
Cluster size	None	0.231 (-0.124, 0.587)	0.202	0.291 (-0.063, 0.644)	0.107
weights	HW	(-0.136, 0.598)	0.217	(-0.074, 0.656)	0.118
	SSDF	(-0.135, 0.597)	0.212	(-0.084, 0.665)	0.126
	SSDF + HW	(-0.146, 0.609)	0.226	(-0.096, 0.677)	0.138
Minimum-	None	0.267 (-0.099, 0.631)	0.153	0.312 (-0.049, 0.673)	0.090
variance	HW	(-0.099, 0.631)	0.153	(-0.055, 0.679)	0.096
weights	SSDF	(-0.109, 0.641)	0.162	(-0.070, 0.694)	0.108
	SSDF + HW	(-0.110, 0.641)	0.162	(-0.077, 0.701)	0.114
Two-level multiple imputation					
No weighting	None	0.264 (-0.137, 0.665)	0.197	0.280 (-0.115, 0.676)	0.164
	HW	(-0.140, 0.667)	0.200	(-0.123, 0.683)	0.173

Continued on next page

Table Continued

		Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p
	SSDF	(-0.141, 0.669)	0.202	(-0.141, 0.669)	0.202
	SSDF + HW	(-0.144, 0.671)	0.204	(-0.135, 0.696)	0.186
Cluster size	None	0.204 (-0.182, 0.591)	0.300	0.244 (-0.138, 0.625)	0.211
weights	HW	(-0.184, 0.593)	0.303	(-0.143, 0.630)	0.217
	SSDF	(-0.186, 0.594)	0.305	(-0.149, 0.637)	0.224
	SSDF + HW	(-0.188, 0.597)	0.308	(-0.155, 0.642)	0.231
Minimum-	None	0.235 (-0.160, 0.630)	0.244	0.260 (-0.129, 0.650)	0.190
variance	HW	(-0.158, 0.628)	0.241	(-0.133, 0.653)	0.194
weights	SSDF	(-0.164, 0.634)	0.248	(-0.141, 0.661)	0.204
	SSDF + HW	(-0.162, 0.632)	0.246	(-0.144, 0.665)	0.207

[a] Adjusted for *PLACE*, *SIZE* and *HOME*.

HW: Huber-White; SSDF: small sample degrees of freedom.

8.2.4 Individual-level analysis

Here, I report the results from ITT and LATE estimations. Since the assessment of residual variance in section 8.2.2 suggested there is variance heterogeneity across trial groups, I first assume homoscedasticity and then allow for level-1 or level-2 variance heterogeneity. The model did not converge when I allow simultaneously for level-1 and level-2 variance heterogeneity.

8.2.4.1 ITT effect

ITT analyses were carried out using mixed effects linear regression on complete records and multilevel multiple imputed datasets. Following the exploratory analysis in section 8.2.2, I tested the adequacy of variance homogeneity assumption, using the likelihood ratio test performed only on the complete records. For the unadjusted ITT analysis, there is strong evidence of level-1 variance heterogeneity (p=0.006) and of level-2 variance heterogeneity (p=0.01) across trial groups. As to the adjusted ITT analysis, no evidence of level-1 and level-2 variance heterogeneity was found (p=0.39 and p=0.75, respectively). This suggests that the omission of individual-level and/or cluster-level covariates is the source of variance heterogeneity. Therefore, the adjusted ITT analysis assuming variance homogeneity seems appropriate.

Table 8.7 shows ITT effect estimates at the individual level. The adjusted ITT effect estimates assuming variance homogeneity are 0.29 (95% CI: -0.06, 0.64; p=0.10) for complete records analysis and 0.23 (95% CI: -0.13, 0.59; p=0.21) for analyses based on multilevel multiple imputation, suggesting no evidence of ITT effect at the individual level. Addressing the potential bias introduced by missing data via multilevel multiple imputation leads to more conservative results. Adjusting for covariates (both cluster and individual levels) provides some efficiency gain. However, unlike CL-ITT analyses, the point estimates of the ITT effect is reduced after covariate adjustment.

Table 8.7: IL-ITT effect estimates (as mean difference) at resident level, on SPPB at 12 months, assuming and relaxing variance homogeneity assumption

	Unadjusted ITT (95% CI)	**p**	**Adjusted ITT (95% CI)[a]**	**p**
Complete Case				
Variance homogeneity	0.342 (-0.128, 0.812)	0.154	0.289 (-0.060, 0.639)	0.104
Level-1 heterogeneity[b]	0.345 (-0.113, 0.804)	0.140	0.290 (-0.060, 0.640)	0.105
Level-2 heterogeneity[b]	0.351 (-0.129, 0.832)	0.152	0.287 (-0.063, 0.637)	0.108
Multilevel multiple imputation				
Variance homogeneity	0.273 (-0.181, 0.727)	0.238	0.231 (-0.126, 0.588)	0.205
Level-1 heterogeneity[b]	0.269 (-0.178, 0.715)	0.238	0.231 (-0.128, 0.590)	0.207
Level-2 heterogeneity[b]	0.289 (-0.176, 0.753)	0.224	0.229 (-0.118, 0.575)	0.195

[a] Adjusted for cluster-level covariates *PLACE*, *SIZE* and *HOME*, and individual-level covariates *AGE*, *SEX*, *ANTIDEP0 MMSE* and *SPPB0*.

[b] Variance heterogeneity across trial groups.

8.2.4.2 Individual-level LATE

Wald and TSLS estimation of LATE at the individual level were carried out in Stata 15. I did not perform the Wald estimation using multiple imputed data because of the complexity of combining bootstrap techniques and multiple imputation. The Bayesian multilevel mixture analysis was implemented in JAGS through R using the "Rjags" package allowing for level-2 variance heterogeneity across either trial groups or adherence classes.

Jeffrey's prior was used for level-1 standard deviation while a Uniform or half-Cauchy

prior for the level-2 standard deviation [114]. I chose a vague normal prior distribution for all model coefficients.

I used 900000 iterations and three chains. The first 100000 iterations were discarded as burn-in. I used such a large number of iterations to ensure good convergence for all model parameters (including variances), and particularly for settings where variance homogeneity was relaxed. The analysis codes in Stata and R are shown in Appendix A.10.

I compare the results allowing for level-2 heterogeneity across trial groups or adherence classes, and level-1 heterogeneity across adherence class only. The results are shown in Table 8.8.

Under variance homogeneity assumption, Wald estimation with bootstrapped SEs, TSLS with Huber-White-Rogers SEs and TSLS with Moulton's correction lead to very similar LATE estimates and 95% CIs. The Wald estimation with bootstrapped SEs appears to lead to slightly conservative results compared to TSLS after covariate adjustment. The individual-level LATE estimates from the unadjusted Bayesian multilevel mixture model are much lower compared to the Wald and TSLS ones for multilevel multiple imputed data. However, after covariate adjustment, the Bayesian multilevel mixture model approach provides LATE estimates close to those obtained by other methods.

The LATE point estimate from the Bayesian model does not substantially change when level-2 variance heterogeneity is assumed across trial groups or adherence classes. However, there is a substantial gain in precision when allowing for level-2 variance heterogeneity across adherence classes instead of trial groups. This precision gain occurs when using a Uniform prior for the level-2 standard deviation, but the LATE estimate reported for half-Cauchy prior did not converge. Results from allowing for level-2 variance heterogeneity across adherence classes suggest weak evidence of causal treatment effect among compliant residents, whereas there was strong evidence of causal treatment effect when level-1 heterogeneity is allowed for.

Except for the poor convergence noted when relaxing level-1 homogeneity across adherence class, the use of a Uniform or half-Cauchy prior for the level-2 standard deviation has little impact on the Bayesian estimation of LATE at the individual level.

Substantive summary

I first conducted ITT analyses using various methods to address the question of whether offering exercise sessions would affect the short physical performance battery neither at the care-home level or resident level. Overall, there was no evidence of the effect of offering those exercise sessions.

Then, I investigated causal analyses by estimating LATE to assess whether actual attendance to exercise sessions would affect the short physical performance battery among the care-homes or residents that would comply to any intervention or condition they would be assigned to. I found no evidence of causal effect of attendance to exercise sessions on the short physical performance battery neither at the care-home level nor resident level in the OPERA trial.

Table 8.8: IL-LATE estimates (as mean difference) at resident level, of attending at least one group exercise session on SPPB at 12 months, assuming ER

	Unadjusted LATE (95% CI)			**Adjusted LATE (95% CI)[a]**		
	Complete Case	**Multilevel MI**	**Multilevel Mixture**	**Complete Case**	**Multilevel MI**	**Multilevel Mixture**
Variance homogeneity						
Wald	0.401 (-0.102, 0.903)	-	-	0.294 (-0.086, 0.675)	-	-
TSLS HWR	0.401 (-0.101, 0.902)	0.295 (-0.214, 0.804)	-	0.294 (-0.068, 0.656)	0.246 (-0.141, 0.634)	-
TSLS Moulton	0.401 (-0.104, 0.905)	0.295 (-0.217, 0.807)	-	0.294 (-0.071, 0.660)	0.246 (-0.147, 0.639)	-
BMM Uniform	-	-	0.152 (-0.195, 0.499)	-	-	0.219 (-0.219, 0.648)
BMM Half-Cauchy	-	-	0.153 (-0.196, 0.500)	-	-	0.218 (-0.219, 0.648)
Level-2 variance heterogeneity across trial groups						
BMM Uniform	-	-	0.437 (-0.116, 0.960)	-	-	0.205 (-0.161, 0.557)
BMM Half-Cauchy	-	-	0.437 (-0.114, 0.961)	-	-	0.205 (-0.162, 0.557)
Level-2 variance heterogeneity across adherence classes						
BMM Uniform	-	-	0.151 (-0.175, 0.479)	-	-	0.206 (-0.047, 0.459)
BMM Half-Cauchy	-	-	0.151 (-0.175, 0.480)	-	-	0.206 (-0.047, 0.459)
Level-1 variance heterogeneity across adherence classes						
BMM Uniform	-	-	0.428 (-0.191, 1.220)	-	-	0.373 (0.010, 0.737)
BMM Half-Cauchy	-	-	0.418 (-0.194, 1.216)	-	-	0.269 (-0.120, 0.801)[b]

[a] Adjusted for cluster-level covariates *PLACE*, *SIZE*, *HOME*, and individual-level covariates *AGE*, *SEX*, *ANTIDEP0*, *MMSE0* and *SPPB0*.

[b] Gelman-Rubin statistics suggests poor mixing of chains.

MI: Multiple imputation. BMM: Bayesian Multilevel Mixture. HWR: Huber-White-Rogers.

8.3 Re-analysis of the TX4FLUJAB trial

This was a CRT of general practices in England aiming at estimating the effect of text messaging influenza vaccine reminders on increasing vaccine uptake in patients with chronic conditions, carried during the 2013 influenza season [36]. General practices (GPs) were stratified by the type of software used for text messaging and randomised to either standard care (control group, 79 GPs and 51136 patients) or a text messaging campaign (active group, 77 GPs and 51121 patients). Practices were not blinded to their allocation. GPs were the unit of analysis and the outcome of interest was the proportion of influenza vaccine uptake at the GP level. Influenza vaccination within the GPs was automatically recorded in the clinical system from which the data were extracted, so there are no missing data.

Since non-adherence was anticipated, the original statistical analysis plan specified obtaining by IV regression an efficacy estimate at the GP level [36]. The original publication reported an estimated increase in vaccine uptake from texting reminders of 14.3% (95% CI –0.59% to 29.2%) [36], after dichotomising adherence at the cluster-level as either 100% of eligible patients, compared with texting $< 100\%$.

Adherence to the intervention at the individual level could not be measured for all practices because it was recorded in a usable form only for GPs using a specific software. Therefore, for these re-analyses, I restrict the dataset to 116 GPs (58 in the intervention and 58 in the standard care arm) for which individual-level adherence data are available. I only show the results for CL-TSLS and the Schochet-Chiang Wald-based estimations.

8.3.1 Descriptive analysis

Six of the 58 practices (10%) in the intervention arm, did not send any reminders. Conversely, 21 of the 58 practices (36% in the standard care arm) actually sent a reminder to at least one patient. Hence non-adherence is two-sided. It also varies at the individual level. The median (range) of percentage of non-adherence at the GP level was 0% (0%-78.4%) and 21.0% (0%-83.5%) in the control and active group, respectively (Table 8.9).

The characteristics of the GPs and of the patients included in these analyses are comparable across trial groups (Table 8.9); further the marginal ICC for individual-level outcome (vaccination) and treatment received (text message reminder) was 0.03 and 0.84 on the log-odds scale, respectively.

Table 8.9: Baseline characteristics and percentages of non-adherence for the TXT4FLUJAB trial

Characteristics	**Control**	**Active**
Practice-level characteristics		
Number of practices, n (%)	58 (100.0)	58 (100.0)
Open on weekends, n (%)	39 (67.2)	37 (63.8)
Patients per practice, median (range)	660 (148-1678)	684 (79-3022)
Patient-level characteristics		
Number of patients, n (%)	40633 (100)	41073 (100)
Male, n (%)	20752 (51.1)	21012 (51.2)
Has any disease, n (%)	39244 (96.6)	39672 (96.6)
Age, median (range)	50 (18-64)	50 (18-64)
Active treatment received		
Patients receiving text message reminders, n (%)	2628 (6.5)	11113 (27.1)
Practices sending text message reminders, n (%)	21 (36.2)	52 (80.7)
% of patients in each GP receiving reminders, median (range)	0 (0-78.4)	21.0 (0-83.5)

8.3.2 Cluster-level analyses

I first discuss the plausibility of the LATE identification assumptions. Then, I present the analysis results.

8.3.2.1 Plausibility of LATE identification assumptions

The unconfoundedness assumption of the CL randomised treatment is satisfied by design. To check whether cluster randomisation is a relevant instrument, I performed a test on the first stage of the CL-TSLS. The corresponding F-statistic is $F(1, 114) = 28.7 > 10$, thus passing Staiger and Stock's rule suggesting that the random treatment assignment is a relevant instrument [106].

The exclusion restriction at the individual level implies that there is no other mechanism by which the GP being randomised to sending text vaccination reminders

can affect a patient's actual vaccination uptake beside via the sending of the message. This assumption needs further justification, as in principle, a GP randomised to send reminders can be more conscious of the risks the patients face during the influenza season and use other means to remind at-risk patients, either in person, by post or by putting out flyers and posters in the clinic. So, it is possible that there are patients who do not receive text reminders and yet are prompted to get vaccinated by other means, by virtue of their practice being in the active group. However, flyers, posters and postal letters already form part of regular care, so we believe they do not really vary by whether the GP is randomised to the active group.

The monotonicity assumption also seems plausible as GPs randomised to the active group were more likely to send a text message reminder than those in the control group (see Table 8.9). Finally, there is a small risk of interference. The cluster defined by GP practice should minimise this, as we only need to assume no interference at the cluster level, but it could be plausible that patients interact with those outside their GP, so that the exposure to a text message reminder of one patient may indeed affect the potential outcome, in this case, influenza vaccination of another patient from a different GP. The risk is small as usually close family members belong to the same general practice.

8.3.2.2 CL-LATE estimation

CL-TSLS and the Schochet-Chiang's method using the unadjusted CL outcome summaries are implemented by adjusting and not adjusting for a baseline CL covariate, namely whether the clinic was open on the weekends (yes/no). Table 8.10 shows the CL-LATE estimates (expressed as mean risk differences), with 95% CIs and p-values obtained via different weighting strategies, and corrections.

Using cluster size weights results in different point estimates from the rest. This was expected as there is substantial cluster size imbalance (cluster size range: 148–1678 in the control group and 79–3022 in the active group (Table 8.9). The results obtained from TSLS using no weights or MV weights leads to point estimates that are very close to those found in the original publication [36]. As expected, the Schochet-

Chiang method and the unweighted TSLS produces the same point estimates. Like in the OPERA trial, the two estimation methods lead to similar conclusions and suggest weak evidence of a positive text messaging effect on the proportion of flu vaccines uptake at the cluster level (p-value between 0.06 and 0.07).

In terms of inference, the use of SSDF correction in calculating CIs is not important, as the number of clusters is large, but the HW SEs paired with MV weighting provides efficiency gains, especially for the adjusted CL-TSLS analyses. Overall however, the CIs are still very wide.

These results suggest that there is weak evidence that receiving a text reminder increases the expected proportion of patients within a compliant practice that get vaccinated against influenza by 14% (95% CI: -0.5 to 30%, $p = 0.065$, based on the adjusted CL-TSLS using MV weights and normal-based CI with HW SEs estimate).

Contrast this with the unadjusted CL-summaries mean risk difference ITT estimate, which indicates a 2.89% increase (95% CI -0.17 to 5.95, $p = 0.064$), highlighting the dilution effects deriving from the non-adherence.

One of the disadvantages of TSLS is lack of efficiency. Adjusting for individual-level baseline covariates may help obtaining narrower CIs. Since CL-TSLS cannot adjust for individual level covariates, we now perform the analyses using "adCL" summary outcomes, generated by adjusting for gender, age and the presence of disease. Results are reported in Table 8.11. The results do not materially change (weak evidence of a 13% increase uptake of vaccination), possibly because these individual-level covariates are not strongly associated with the outcome.

Our illustrative example is limited by the availability of baseline cluster-level variables. Since there was only one CL-variable recorded, the impact of covariate adjustment on the CL-TSLS is negligible. Other limitations of these results include the possibility of measurement error, for if patients received their influenza vaccine outside the practice, this would not have been recorded in the system, unless the patient informed their GP.

Table 8.10: Schochet-Chiang and TSLS estimation of practice-level LATE of reminder text messaging to receive flu vaccine on the percentage uptake of flu vaccine in the TXT4FLUJAB trial using unadjusted CL outcomes, adjusting for individual-level covariates gender, age and presence of disease

		Unadjusted		Adjusted[a]	
		LATE (95% CI)	**p**	**LATE (95% CI)**	**p**
Schochet TSLS	None	0.149 (-0.008, 0.306)	0.063	0.148 (-0.010, 0.306)	0.066
No weighting	None	0.149 (-0.006,0.305)	0.060	0.148 (-0.078,0.303)	0.063
	HW	(-0.006,0.305)	0.060	(-0.005,0.301)	0.058
	SSDF	(-0.009,0.308)	0.065	(-0.012,0.308)	0.069
	SSDF + HW	(-0.009,0.308)	0.065	(-0.009,0.305)	0.064
Cluster size weights	None	0.071 (-0.065,0.207)	0.307	0.074 (-0.061,0.209)	0.284
	HW	(-0.088,0.230)	0.382	(-0.077,0.225)	0.338
	SSDF	(-0.068,0.209)	0.313	(-0.064,0.212)	0.292
	SSDF + HW	(-0.091,0.233)	0.388	(-0.081,0.228)	0.346
Minimum-variance weights	None	0.143 (-0.008,0.293)	0.064	0.142 (-0.009,0.293)	0.065
	HW	(-0.006,0.291)	0.060	(-0.005,0.289)	0.058
	SSDF	(-0.011,0.296)	0.069	(-0.012,0.297)	0.071
	SSDF + HW	(-0.009,0.294)	0.065	(-0.008,0.293)	0.064

[a] Adjusted for whether clinic is opened during weekends.

HW: Huber-White; SSDF: small sample degrees of freedom.

Table 8.11: TSLS estimation of practice-level LATE of reminder text messaging to receive flu vaccine on the percentage uptake of flu vaccine in the TXT4FLUJAB trial using adjusted CL outcomes, adjusting for individual-level covariates gender, age and presence of disease

		Unadjusted		Adjusted[a]	
		LATE (95% CI)	**p**	**LATE (95% CI)**	**p**
No weighting	None	0.133 (-0.016,0.282)	0.081	0.133 (-0.017,0.282)	0.082
	HW	(-0.016,0.282)	0.081	(-0.014,0.280)	0.077
	SSDF	(-0.019,0.285)	0.086	(-0.021,0.286)	0.089
	SSDF + HW	(-0.019,0.285)	0.086	(-0.018,0.283)	0.083
Cluster size weighting	None	0.068 (-0.063,0.198)	0.310	0.071 (-0.058,0.200)	0.280
	Huber-White	(-0.081,0.216)	0.372	(-0.069,0.212)	0.320
	SSDF	(-0.065,0.201)	0.316	(-0.061,0.203)	0.288
	SSDF + HW	(-0.084,0.219)	0.378	(-0.073,0.215)	0.328
Minimum-variance weighting	None	0.128 (-0.017,0.273)	0.084	0.128 (-0.017,0.273)	0.084
	HW	(-0.015,0.271)	0.080	(-0.014,0.269)	0.077
	SSDF	(-0.020,0.275)	0.090	(-0.021,0.277)	0.091
	SSDF + HW	(-0.018,0.273)	0.086	(-0.017,0.273)	0.083

[a] TSLS estimation was adjusted for weekend clinics (yes/no).

HW: Huber-White; SSDF: small sample degrees of freedom.

8.3.3 Individual-level analyses

Table 8.12 shows the results from the Wald estimation and TSLS with HWR or Moulton correction, with and without covariate adjustment. For the Wald estimation, I obtained the 95% CI for LATE via bootstrapping using 1500 replicates and normal approximation. In order to get the risk difference from the Bayesian multilevel mixture modelling, I marginalize the individual-level probability of receiving influenza vaccine over the space of random effects using the Monte Carlo integration, for the control and active groups separately. Unfortunately, I encountered a convergence issue potentially due to the large number of clusters and individuals and therefore cannot present any results for this analysis. The Wald and TSLS (with HWR and Moulton correction) estimations lead to very similar results and suggest no evidence of text messaging on influenza vaccine uptake.

Substantive summary

I conducted causal analyses by estimating LATE using several methods to address the question of whether actually receiving a text messaging reminder of influenza vaccine would affect the uptake of influenza vaccine among GPs or patients that would comply to any condition they would be randomised to. I found weak evidence of positive causal effect of text messaging reminder of influenza vaccination on the uptake of influenza vaccine at the GP level but no evidence of causal effect at the patient level in the TXT4FLUJAB trial.

Table 8.12: IL-LATE estimates (as mean difference) at patient level, of text message reminders to receive flu vaccination on the uptake of flu vaccine in the TXT4FLUJAB trial, assuming and relaxing variance homogeneity assumption and adjusting/not adjusting for gender, age, presence of disease and whether clinic is opened during weekends

	Unadjusted		**Adjusted**	
	LATE (95% CI)	**p**	**LATE (95% CI)[a]**	**p**
TSLS HWR	0.071 (-0.088, 0.230)	0.382	0.071 (-0.069, 0.211)	0.320
TSLS Moulton	0.071 (-0.086, 0.228)	0.377	0.071 (-0.076, 0.218)	0.342
Wald ratio [b]	0.071 (-0.093, 0.235)	0.398	0.071 (-0.075, 0.217)	0.338

[a] Adjusted for cluster-level covariate "whether clinic is opened during weekends" and individual-level gender, age and presence of disease.

[b] 95% CI obtained bootstrap-based normal approximation, with 1500 replicates.

HWR: Huber-White-Rogers.

8.4 Summary

Covariate adjustment offers some efficiency gain whether CL-summaries or individual-level analyses are performed particularly when the covariates are associated with the outcome, with or without weighting using either complete records or multiple imputed data.

As expected, the Schochet-Chiang method and unweighted CL-TSLS provide the same CL-LATE point estimates. Moreover, estimates from the Schochet-Chiang method and unweighted CL-TSLS result in similar conclusions. The relatively large number of clusters makes the between-cluster variance likely reliable and therefore, the CL-TSLS with MV weighting seems appropriate to report. As for IL-LATE, the Wald estimation with bootstrapped SEs and TSLS (with HWR or Moulton correction) estimation also lead to very similar results.

The Bayesian multilevel mixture is suitable when the assumption of variance homogeneity across trial groups is to be relaxed. Relaxing the variance homogeneity assumption either at the cluster level or at the individual level may favour efficiency. A larger number of iterations may be required when allowing for level-1 variance heterogeneity and using a half-Cauchy prior for level-2 standard deviation. These findings provide additional insights as I did not investigate via simulations the performance of the Bayesian multilevel mixture under variance heterogeneity scenario.

In applying these methods to the analyses of two CRTs, I have investigated whether the actual uptake of the randomised intervention among compliers had a causal effect on the outcomes of interest. I found – (i) weak evidence of positive causal effect of text messaging reminder of influenza vaccination on the uptake of influenza vaccine at the care-home level but no evidence of causal effect at the resident level in the TXT4FLUJAB trial and – (ii) no evidence of causal effect of attendance to exercise sessions on the short physical performance battery neither at the care-home level nor resident level in the OPERA trial. In interpreting the results, I have also considered the plausibility of the relevant assumptions. Given the design and implementation of these two trials, ER and *monotonicity* are justifiable assumptions. With regards to

the other assumptions, I have examined their plausibility by relaxing some of them (variance homogeneity at each level across trial groups or adherence classes) and found that there is a substantial gain in efficiency when allowing for level-2 variance heterogeneity across adherence classes instead of trial groups in the OPERA trial.

Chapter 9.

Illustration of sensitivity analyses using the OPERA trial data

9.1 Introduction

The present chapter is an extension to chapter 8 and explores some sensitivity analyses to ascertain the robustness of the estimates of IL-LATE in the OPERA trial. Among the IV assumptions (defined in chapter 4), those of *unconfoundedness at cluster level* and *instrument relevance* are often met in one-sided non-adherence trials. Furthermore, in one-sided non-adherence trials, the *monotonicity* assumption is met by design. However, departures from the ER assumption may occur especially in psychological interventions like the OPERA trial, where some participants in the control group may get more depressed for not being offered the active treatment. The purpose of the sensitivity analyses is to assess how robust the results are if there are departures from key untestable assumptions or how severe violations of those assumptions must be to reverse the results.

Therefore, I hypothesize that residents whose care-homes are randomised to the control group in the OPERA trial would be demotivated during the assessment of their functional mobility, resulting in lower SPBB scores at 12 months. In other words, not being offered the active treatment would negatively affect their motivation to do the required exercises for SPPB assessment. I investigated the impact of relaxing the ER assumption when implementing TSLS and Bayesian multilevel mixture methods for estimating individual-level LATE.

The chapter is structured as follows. Section 9.2 introduces the Baiocchi and Conley sensitivity analysis approaches. Section 9.3 presents the results of applying these sensitivity analyses to both TSLS and Bayesian multilevel mixture estimation. Fi-

nally, section 9.4 summarizes the chapter.

9.2 Sensitivity analysis approaches

I first introduce the sensitivity analysis approach for TSLS estimation as in Baiocchi *et al.* [35] which is a special type of relaxing ER assumption only in the exposed. Then, I present the Conley's approach for the Bayesian mixture [33] to cope with more general types of direct effects of treatment assignment. Using the same ideas as in Baiocchi *et al.* [35], I elicited the sensitivity parameter, and used this within the Conley's method so the prior was centred around the sensitivity parameter. The codes for sensitivity analyses in Stata (for TSLS) and R (for Bayesian multilevel mixture) can be found in Appendix A.12.

9.2.1 TSLS estimation

Angrist *et al.* [24] pointed out that the ER assumption does not hold if the IV has an effect on the outcome while holding the value of the treatment received fixed. Recall that the ER assumption is violated when there is a direct effect of treatment assignment (*i.e.* an association not mediated by the treatment received) on outcome. Using data from an observational study, Baiocchi *et al.* [35] relaxed the ER assumption by allowing for a pre-specified interaction effect between the IV and treatment received on the outcome. They used TSLS method so that the original outcome is replaced by a new variable, which I refer to as ER-"adjusted" outcome (adjusted by the effect of the IV that is not mediated by treatment received). This "adjusted" outcome is generated as the difference between the original outcome and the fitted interaction between the IV and the treatment received.

The interaction effect used by Baiocchi [35] as a sensitivity parameter is a particular type of effect of the IV on Y, where there is an added effect on top of the effect of the treatment received D alone among the exposed. This interaction effect has to be posited from external knowledge in order to point-identify LATE when relaxing the ER. Denote by λ the hypothetical added effect of treatment received on the observed outcome Y_{ij} among the exposed and Y_{ij}^{adj} the "adjusted" outcome. The outcome equation that allows for an IV-treatment received interaction and the

adjusted outcome are formulated by Baiocchi *et al.* [35] as follows.

$$Y_{ij} = \beta_0 + \beta D_{ij} + \beta_{\mathrm{W}} W_j + \beta_{\mathrm{X}} X_{ij} + \lambda(1 - Z_j)D_{ij} + \epsilon_{2_{ij}} \tag{9.1}$$

$$Y_{ij}^{adj} = Y_{ij} - \lambda(1 - Z_j)D_{ij} \tag{9.2}$$

where $\epsilon_{2_{ij}} \sim N\left(0, \sigma^2_{\epsilon_2}\right)$. The term $\lambda(1 - Z_j)D_{ij}$ expresses the ER violation on the exposed only. However, none of the residents in the control groups received the treatment. Thus, equation 9.1 is modified to allow for direct Z effect on both exposed and unexposed as shown in equation 9.3.

$$Y_{ij} = \beta_0 + \beta D_{ij} + \beta_{\mathrm{W}} W_j + \beta_{\mathrm{X}} X_{ij} + \lambda(1 - Z_j)(1 - D_{ij}) + \epsilon_{2_{ij}} \tag{9.3}$$

Then, the "adjusted" outcome is given by

$$Y_{ij}^{adj} = Y_{ij} - \lambda(1 - Z_j)(1 - D_{ij}) \tag{9.4}$$

I allow for two hypothetical direct effects of treatment assignment on the outcome *SPPB2* by postulating that residents whose care-homes are assigned to the control group have their *SPPB* at 12 months reduced either by 5% (low effect) or by 30% (high effect) of the ITT estimates. This enables us to ascertain how severe departures from the ER must be to reverse the conclusions regarding the causal treatment effect.

9.2.2 Bayesian multilevel mixture model

Conley *et al.* [33] proposed a Bayesian estimation using non-clustered data from an observational study where the IV is assumed to be "plausibly exogenous", that is, the ER assumption is weakly relaxed such that the direct effect of the IV on the outcome is allowed to be local-to-zero (*i.e.* close to 0) rather than exactly zero. As per Conley [33], I assume a local-to-zero prior for the direct effect of treatment assignment on *SPPB* at 12 months but also make alternative assumptions. The assumptions investigated to assess departures from the ER assumption are as follows. I assume a normal distribution for the direct effect of treatment assignment either – (i) with mean 0 and a large precision τ (here, I use $\tau = 1000$) which is a local-to-0 prior, – (ii) with mean λ different from zero and a large precision τ (here, $\tau = 1000$

as well) and finally – (iii) an uninformative prior with mean 0 and a precision of $\tau = 0.001$. I consider both Uniform and half-Cauchy prior distributions for the standard deviation of the random effects as recommended [114].

I extend Conley's approach [33] to one-sided non-adherence CRTs using the following model.

$$\begin{cases} Y_{ij} \sim N(\mu_{ij_{\mathrm{C}}}, \sigma^2_{\upsilon_{\mathrm{C}}} + \sigma^2_{\epsilon_{\mathrm{C}}}) \; ; C \in \{0,1\} \\ \mu_{ij_0} = \beta_{0,0} + \beta_{\mathrm{Z},0} Z_j + \beta_{\mathrm{W},0} W_j + \beta_{\mathrm{X},0} X_{ij} + \upsilon_{j_0} \\ \mu_{ij_1} = \beta_{0,1} + \beta_{\mathrm{Z},1} Z_j + \beta_{\mathrm{W},1} W_j + \beta_{\mathrm{X},1} X_{ij} + \upsilon_{j_1} \\ C_{ij} \sim \mathrm{Bern}(p_{ij}) \\ \mathrm{logit}(p_{ij}) = \lambda_0 + \lambda_W W_j + \lambda_X X_{ij} + \zeta_j \end{cases} \tag{9.5}$$

where C is the latent adherence class such that $C_{ij} = D_{ij}$ if $Z_j = 1$ and $C_{ij} =$ missing if $Z_j = 0$. Here $1 =$ "complier" and $0 =$ "never-taker". $\upsilon^0_j \sim N(0, \sigma^2_{\upsilon^0})$, $\upsilon^1_j \sim N(0, \sigma^2_{\upsilon^1})$ and $\zeta_j \sim N(0, \sigma^2_{\zeta})$.

The ER assumption is relaxed by formulating priors for $\beta_{\mathrm{Z},0}$ in equation (9.5) as follows: – (i) $\beta_{\mathrm{Z},0} \sim N(\lambda, \tau = 1000)$, – (ii) $\beta_{\mathrm{Z},0} \sim N(\lambda, \tau = 1000)$ and – (iii) $\beta_{\mathrm{Z},0} \sim N(0, \tau = 0.001)$. Like in section 9.2.1, λ is either 5% or 30% of the ITT effects.

9.3 Results

Table 9.1 shows the IL-LATE estimates obtained assuming and relaxing the ER assumption. Estimations are performed using TSLS with Huber-White-Rogers or Moulton SEs and the Bayesian multilevel mixture model. I assume level-1 and level-2 variance homogeneity. Table 9.2 shows the IL-LATE estimates obtained from fitting the Bayesian multilevel mixture model only, assuming and relaxing ER assumption but with level-2 variance heterogeneity across trial groups or adherence classes.

TSLS estimation of IL-LATE is sensitive to the size of the pre-specified direct effect of treatment assignment on outcome, regardless of covariate adjustment and the SEs estimation approach (HWR or Moulton correction). Although the IL-LATE point estimates have substantially changed and the 95% CIs get narrower after adjusting for covariates, there is no noticeable gain in efficiency. The p-values,

whether adjusted for covariates or not, seem to increase with larger sizes of the direct effect of treatment assignment on outcome. This may suggest that TSLS estimation may be more inefficient when relaxing ER and this inefficiency gets worse as the size of the direct effect of treatment assignment increases. As noted under the ER assumption, estimating the SEs using HWR or Moulton's correction leads to very similar results.

Unlike TSLS, the IL-LATE estimates from the Bayesian multilevel mixture model seem in general less affected by the assumption made on the direct effect of treatment assignment on outcome. However, when allowing for level-2 variance heterogeneity across trial groups, using an uninformative prior with a mean 0 for the direct effect of treatment assignment without covariate adjustment results in IL-LATE estimates with opposite signs compared to those obtained from other analysis scenarios. In addition, the resulting 95% credible intervals do not overlap with the rest of the estimates. This is corrected after adjusting for covariates, leading to conclusions that are similar across all the assumptions made on the sensitivity parameter. For the scenario where I allow for level-2 variance heterogeneity across trial groups and no covariates are adjusted for, the posterior distribution of the treatment effect in never-takers when using a vague prior (mean 0 with very small precision) results in large values (median: 5.97; 2.5^{th} to 97.5^{th} : 4.95 to 7.02) in contrast with the small values (median: -0.0002; 2.5^{th} to 97.5^{th} : -0.06 to 0.06) obtained when a local-to-0 prior for the treatment effect in never-takers is used. Allowing for variance heterogeneity across trial groups coupled with no covariate adjustment may result in poor estimates of the variance in each trial group because of the reduced number of clusters contributing to the estimation, as opposed to allowing for the heterogeneity across adherence classes where all the clusters potentially contribute to estimating the outcome variance. The fact that IL-LATE estimates are downward when using a vague prior may be because the adherence classes are badly predicted as no covariates are included in the model coupled with the weak identification of LATE subsequent to a strong violation of the ER assumption, which translates via the vague prior on the treatment effect in never-takers. The fact that IL-LATE estimates are similar

across all the sensitivity assumptions after covariate adjustment supports the claim of poor prediction of adherence classes in the absence of covariates, as the covariates have information that helps predict the adherence classes.

The prior distribution used for the level-2 standard deviations, whether Uniform or Half-Cauchy, has little impact on the results. IL-LATE estimation using Bayesian multilevel mixture modelling appears to be more efficient than TSLS.

When estimation is done using Bayesian multilevel mixture model, efficiency is gained by allowing for level-2 variance heterogeneity across adherence classes rather than trial groups, regardless of the assumptions made on the size of the direct effect of treatment assignment on outcome and whether covariates are included in the model or not.

As to the suitability of relaxing ER, the results from the Bayesian multilevel mixture modelling support the plausibility of ER assumption in the OPERA trial as the LATE estimates assuming or relaxing ER are very similar with or without covariate adjustment and irrespective of the assumption made on the direct effect of treatment assignment on outcome. The results validate our assumption of direct treatment assignment effect and suggest that only attendance to at least one exercise session may have an effect on *SPPB* at 12 months and there is no noticeable effect of treatment allocation. The plausibility of the ER assumption may be due to the "all-or-none" (attending at least one exercise session or not at all) working definition of adherence to treatment adopted in this thesis. This definition of adherence is less susceptible to violation of the ER assumption as even little exposure to exercise sessions may induced a non-zero effect on the outcome than no exposure at all may do.

Moreover, when allowing for level-2 variance heterogeneity across adherence classes, the results are suggestive of weak evidence of causal effects of exercise sessions on SPPB at 12 months. More precisely, attending at least one exercise session increases SPPB at 12 months by approximately 0.21 on average and there is 95% probability that this effect is roughly within -0.05 to 0.46 in residents who would attend at least

one exercise session had they been assigned to the active group but would not attend any exercise session at all if they were assigned to the control group.

Table 9.1: Individual-level LATE estimates expressed as a mean difference on SPPB at 12 months with/without the exclusion-restriction assumption and assuming variance homogeneity, adjusting and not adjusting for covariates and obtained on complete records and multiple imputed data

		Unadjusted		**Adjusted[b]**	
		LATE (95% CI)[a]	**p**	**LATE (95% CI)**	**p**
Complete records					
TSLS	Assumed ER	0.401 (-0.101, 0.902)	0.117	0.294 (-0.068, 0.656)	0.111
HWR	$\lambda = 5\%$ ITT	0.382 (-0.120, 0.884)	0.135	0.279 (-0.084, 0.641)	0.132
	$\lambda = 30\%$ ITT	0.290 (-0.213, 0.793)	0.258	0.201 (-0.163, 0.564)	0.279
TSLS	Assumed ER	0.401 (-0.104, 0.905)	0.119	0.294 (-0.071, 0.905)	0.115
Moulton	$\lambda = 5\%$ ITT	0.382 (-0.122, 0.887)	0.138	0.279 (-0.087, 0.644)	0.135
	$\lambda = 30\%$ ITT	0.290 (-0.215, 0.795)	0.261	0.201 (-0.165, 0.567)	0.282
Bayesian	Assumed ER	0.152 (-0.195, 0.499)	-	0.219 (-0.219, 0.648)	-
Multilevel	Local-to-0	0.153 (-0.194, 0.499)	-	0.211 (-0.230, 0.644)	-
Mixture,	$\lambda = 0$, vague prior	0.266 (-0.087, 0.619)	-	0.195 (-0.181, 0.564)	-
Uniform	$\lambda = 5\%$ ITT	0.151 (-0.196, 0.498)	-	0.212 (-0.232, 0.645)	-
	$\lambda = 30\%$ ITT	0.146 (-0.202, 0.492)	-	0.213 (-0.232, 0.652)	-
Bayesian	Assumed ER	0.153 (-0.196, 0.500)	-	0.218 (-0.219, 0.648)	-
Multilevel	Local-to-0	0.153 (-0.196, 0.500)	-	0.211 (-0.230, 0.643)	-
Mixture,	$\lambda = 0$, vague prior	0.266 (-0.087, 0.620)	-	0.195 (-0.184, 0.566)	-
Half-Cauchy	$\lambda = 5\%$ ITT	0.152 (-0.197, 0.500)	-	0.211 (-0.230, 0.645)	-
	$\lambda = 30\%$ ITT	0.147 (-0.201, 0.494)	-	0.214 (-0.228, 0.647)	-
Multilevel multiple imputation					
TSLS	Assumed ER	0.295 (-0.214, 0.804)	0.256	0.246 (-0.141, 0.634)	0.213
HWR	$\lambda = 5\%$ ITT	0.303 (-0.229, 0.836)	0.264	0.229 (-0.141, 0.598)	0.225
	$\lambda = 30\%$ ITT	0.303 (-0.229, 0.836)	0.264	0.229 (-0.141, 0.598)	0.226
TSLS	Assumed ER	0.295 (-0.217, 0.807)	0.258	0.246 (-0.147, 0.639)	0.219
Moulton	$\lambda = 5\%$ ITT	0.272 (-0.248, 0.791)	0.306	0.220 (-0.177, 0.617)	0.276
	$\lambda = 30\%$ ITT	0.272 (-0.248, 0.791)	0.306	0.220 (-0.177, 0.617)	0.276

[a] Confidence intervals for estimates from TSLS and credible intervals for the Bayesian models.

[b] Adjusted for cluster-level covariates *PLACE*, *SIZE* and *HOME*, and individual-level covariates *AGE*, *SEX*, *ANTIDEP0 MMSE* and *SPPB0*.

HWR: Huber-White-Roger. ER: exclusion-restriction. Uniform or Half-Cauchy prior for the level-2 standard deviations.

Table 9.2: Individual-level LATE estimates expressed as a mean difference on SPPB at 12 months with/without the exclusion-restriction assumption and assuming level-2 variance heterogeneity across trial groups or adherence classes, adjusting and not adjusting for covariates

		Unadjusted LATE (95% CI)[a]	**Adjusted[b] LATE (95% CI)**
Level-2 variance heterogeneity across trial groups			
Bayesian	Assumed ER	0.437 (-0.116, 0.960)	0.205 (-0.219, 0.557)
Multilevel	Local-to-0	0.467 (-0.113, 0.960)	0.225 (-0.147, 0.584)
Mixture,	$\lambda = 0$, vague prior	-0.197 (-0.621, 0.222)	0.195 (-0.128, 0.519)
Uniform	$\lambda = 5\%$ ITT	0.439 (-0.110, 0.960)	0.224 (-0.148, 0.584)
	$\lambda = 30\%$ ITT	0.442 (-0.103, 0.960)	0.223 (-0.152, 0.582)
Bayesian	Assumed ER	0.437 (-0.114, 0.961)	0.205 (-0.162, 0.557)
Multilevel	Local-to-0	0.437 (-0.111, 0.961)	0.196 (-0.172, 0.547)
Mixture,	$\lambda = 0$, vague prior	-0.197 (-0.622, 0.224)	0.196 (-0.127, 0.519)
Half-Cauchy	$\lambda = 5\%$ ITT	0.438 (-0.111, 0.960)	0.196 (-0.172, 0.547)
	$\lambda = 30\%$ ITT	0.442 (-0.104, 0.961)	0.193 (-0.178, 0.546)
Level-2 variance heterogeneity across adherence classes			
Bayesian	Assumed ER	0.151 (-0.175, 0.479)	0.206 (-0.047, 0.459)
Multilevel	Local-to-0	0.151 (-0.175, 0.480)	0.204 (-0.050, 0.457)
Mixture,	$\lambda = 0$, vague prior	0.188 (-0.144, 0.524)	0.200 (-0.054, 0.453)
Uniform	$\lambda = 5\%$ ITT	0.151 (-0.174, 0.479)	0.223 (-0.052, 0.584)
	$\lambda = 30\%$ ITT	0.149 (-0.178, 0.477)	0.204 (-0.049, 0.457)
Bayesian	Assumed ER	0.151 (-0.175, 0.480)	0.206 (-0.047, 0.459)
Multilevel	Local-to-0	0.151 (-0.174, 0.480)	0.204 (-0.050, 0.458)
Mixture,	$\lambda = 0$, vague prior	0.188 (-0.145, 0.525)	0.200 (-0.054, 0.453)
Half-Cauchy	$\lambda = 5\%$ ITT	0.150 (-0.175, 0.479)	0.204 (-0.050, 0.458)
	$\lambda = 30\%$ ITT	0.150 (-0.176, 0.477)	0.225 (-0.052, 0.547)

[a] Confidence intervals for estimates from TSLS and credible intervals for the Bayesian models.

[b] Adjusted for cluster-level covariates *PLACE*, *SIZE* and *HOME*, and individual-level covariates *AGE*, *SEX*, *ANTIDEP0 MMSE* and *SPPB0*.

HWR: Huber-White-Roger. ER: exclusion-restriction. Uniform or Half-Cauchy prior for the level-2 standard deviations.

9.4 Summary

TSLS estimation of LATE when applied to the OPERA trial is sensitive to the pre-specified size of the direct effect of treatment assignment on outcome. For that same trial, I note that TSLS estimation appears to be inefficient and the inefficiency degenerates with increasing size of the direct effect of treatment assignment on outcome. On the contrary, the LATE estimates from the Bayesian multilevel mixture modelling are not sensitive to the assumptions made on the size of the direct effect of treatment assignment on outcome. Estimation using the Bayesian multilevel mixture modelling appears to be more efficient compared to TSLS. Efficiency is gained in the OPERA trial after allowing for level-2 heterogeneity across adherence class rather across trial groups. Based on the Bayesian multilevel mixture modelling which offers a more elaborate approach for sensitivity analyses, assuming ER in the OPERA trial is plausible. However, those contradictory findings between TSLS and the Bayesian multilevel mixture modelling are specific to the OPERA trial and therefore may not be generalised to all CRTs.

Chapter 10.

Discussion

This final chapter summarises and discusses key findings along with some recommendations to guide trial investigators interested in estimating LATE at the cluster or individual level in CRTs where there is non-adherence. The chapter is structured as follows. Section 10.1 highlights key findings from the systematic review on the reporting of non-adherence and how it is addressed (chapter 2) and from the estimation of cluster-level LATE (chapters 4, 5 and 8) and individual-level LATE (chapters 6, 7, 8 and 9). Then, I point out the strengths and limitations in section 10.2, and I outline the practical implications of the thesis in section 10.3 and future avenues of research on the estimation of causal treatment effects in section 10.4. I end with a conclusion in section 10.5.

10.1 Summary of findings

This thesis contributes to the literature on the estimation of causal treatment effects in CRTs where there is non-adherence to treatment and focuses on IV-based methods which enable the estimation of causal effects even in the presence of unmeasured confounders [27]. I first ascertained the current practice of reporting and addressing non-adherence when causal treatment effects are of interest in CRTs via a systematic review. Then, I introduced and assessed the performance of IV-based methods for estimating cluster-level LATE and individual-level LATE through simulations under the required identification assumptions for LATE. In addition, I performed some sensitivity analyses for individual-level LATE estimation particularly. Two motivating examples of CRTs, namely the OPERA [37] and TXT4FLUJAB [36] trials, have been used to illustrate those methods.

From the systematic review which includes CRT reports published in 2011 and may be outdated, I estimated the prevalence of non-adherence and reported the

methods used to estimate causal treatment effects. This has to be interpreted in light of the fact that there has been many recent methodological developments and an increased awareness of the need for causal inference methods for clinical trials with non-adherence. In particular, the 2017 addendum to the ICH E9 guidelines [45] has focused attention on LATE estimands. Therefore, the conclusions drawn in my review, which pre-dated this publication, may no longer apply to the current reporting and analysis practice.

About half of the studies included information on treatment adherence. The non-reporting on treatment adherence by the remaining half of the studies may be because of a poorer methodological quality than the other half of the studies that provided information on adherence, so-called the "exclusion paradox" [67, 68] and may have encountered non-adherence to treatment. Non-adherence to treatment was common in CRTs but it was not sufficiently well reported. Cluster-level non-adherence was less common than individual-level non-adherence. Moreover, after taking into consideration possible under-reporting, the overall level of non-adherence in CRTs may be comparable with that in RCTs reported in Dodd et al. [11], contradicting the claim that a CRT design improves adherence to treatment. All of the reviewed CRTs that reported adherence-adjusted estimates performed easy-to-implement methods such as per-protocol and as-treated, without discussing the plausibility of the very strong assumptions necessary for such analyses to result in unbiased causal treatment estimates [18, 19]. No study estimated LATE or any other appropriate statistical methods under more plausible assumptions for unbiased causal estimation [24, 34]. More effort should be made to improve the quality of reporting of non-adherence.

I proposed some recommendations for CRT investigators which are summarised in Box 10.1 and published in Clinical Trials (see appendix A.3). Previous recommendations for reporting adherence and conducting causal analyses for RCTs are still relevant [11], and I encourage researchers to follow these as much as possible.

Box 10.1: Guidelines for analysing and reporting cluster randomised trials with non-adherence to treatment

1. Report how adherence to treatment is defined and measured. Describe adherence at the cluster and individual level. If dichotomised, justify the choice of threshold made. These choices should be pre-specified in the protocol [11].
2. Where there is interest in the causal treatment effect, this should be stated clearly in the trial protocol, prior to data collection.
3. Adherence measures should be collected alongside other trial data.
4. Report the number of clusters and individuals that received the intended treatment in each trial arm [39].
5. Details of the planned causal analyses should be included in the statistical analysis plan, in advance of receiving the data.
6. Efforts should be made in the statistical analysis to reduce any bias introduced by the fact that treatment received may be associated with other variables affecting the outcome.
7. Choose a statistical method that relies on assumptions that are clarified for the estimates to be valid and interpret non-adherence adjusted analyses as explanatory.
8. Discuss the assumptions necessary for the chosen analysis method to result in unbiased causal treatment effect estimates and their plausibility in the context of the CRT being analysed and reported.
9. In particular, the use of per protocol analysis must be supported by an explanation of why it is reasonable to assume that the group of participants and clusters who did and did not deviate from their allocated treatment are equivalent.
10. If clusters or individuals are excluded from analyses, describe if the fraction excluded is similar between arms, and that the included groups were comparable at baseline [18].
11. Use a method that accounts for clustering adequately. Principal stratification can be used to estimate the LATE while accounting for clustering; alternatives include multilevel mixture models [30] and Bayesian hierarchical models [34]. Alternatively, IV methods can use sandwich variance estimation, which is robust to clustering [35].
12. Sensitivity analyses should be considered when the assumptions necessary for the primary causal analysis are likely to be violated [34, 35].
13. A discussion of potential bias introduced by assumptions' violations in any of the causal analyses should be included in the published report.

It is also important to remember that for CRTs, the validity of the results also relies on obtaining an appropriate estimate for the standard errors, for which it

is crucial to use a method that correctly models the dependence structure of the data [78]. Causal methods that accommodate the clustering should be made more widely available and easy to implement in commonly used software. To promote their use in practical applications, there is need for more tutorial papers describing clearly the assumptions needed and detailing the challenges of performing such adherence-adjusted analyses in the context of a good empirical example. To this end, I investigated the performance of methods estimating LATE in CRTs, in terms of empirical bias and coverage of the 95% CI. Information from the systematic review such as the median proportion of non-adherence at the cluster or individual level, the median ICC for the outcome and the median number of randomised clusters per arm and of number of individuals guided in simulating CRTs.

I first conducted extensive simulations to assess how TSLS using cluster-level outcome summaries and the Schochet-Chiang approach perform when estimating LATE at the cluster level. The simulation study suggests that CL-TSLS outperforms the Schochet-Chiang method in some settings. The Schochet-Chiang method has good coverage when the number of clusters is large irrespective of the settings or when no covariate adjustment is done except for the settings with low ICC for Y and large LATE size. For those settings, the performance of the Schochet-Chiang and TSLS methods are comparable. However, the TSLS estimator with at least SSDF correction is preferable to the Schochet-Chiang method especially for settings where the number of clusters is small. For CL-TSLS, all weighting strategies performed similarly when the number of clusters is not small. When the number of clusters is small, minimum-variance weights tend to be badly estimated and therefore are not recommended. Furthermore, when the cluster sizes are very variable, cluster size weights should not be used. Although in the simulations the choice of weights did not affect the point estimates, these were affected in the illustrative example. Overall the results show that, unless there are very few clusters, or the outcome ICC is large, minimum-variance weighting performs well [74].

Based on the simulations study, Table 10.1 summarises when to adjust for confounders, to use weighted least square and to correct for degrees of freedom in small

samples while performing CL-TSLS. This summary of findings has been published in Statistical Methods in Medical Research (see Appendix A.4).

Table 10.1: Summary of how to perform CL-TSLS

Adherence	**Comments**
At CL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	Adjust for CL variables in TSLS to reduce bias and improve efficiency
If an IL variable is a strong confounder	Use adjusted CL-outcomes in the TSLS to improve efficiency
If CS are imbalanced	Use small sample DF correction to improve inference
At IL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	Avoid adjusting for CL variables
If CS are imbalanced	Use small sample DF correction to improve inference and avoid using CS weights

CS: cluster sizes; CL: cluster level; DF: degrees of freedom; IL: individual level

CL-TSLS is attractive and easy to implement, but it suffers from being inefficient. Cluster-level summary analyses are in general inefficient unless the cluster sizes are (almost) equal [128] and TSLS is also known to be inefficient. However, adjusting for covariates can improve the efficiency [99]. In the context of cluster-level summary analyses, it is only possible to include cluster-level covariates in the regressions [76]. However, I tested the performance of cluster-level outcome summaries which are adjusted for individual-level covariates [8], and showed that this indeed has the potential to improve efficiency in certain settings.

For CL-TSLS analyses, inference should be based on the number of clusters, with CIs constructed by using t-distributions with degrees of freedom equal to $J - p$ [129], where J is the number of clusters and p the number of parameters in TSLS regression. The outcome ICC value is important too, with higher ICCs requiring a larger number of clusters for the asymptotical arguments to work, as well as whether the cluster-level variances are homoscedastic [76].

In view of existing literature and my findings, I conclude with some recommendations for trial analysts when cluster-level LATE is of interest. Those recommendations are summarised in Box 10.1 below.

Box 10.1: Recommendations for cluster-level LATE estimation in CRTs where there is non-adherence to treatment

Which cluster-level summary approach to choose?

1. For simplicity, use unadjusted cluster-level outcome summaries if covariate adjustment is not planned for in the trial protocol.
2. If covariate adjustment is planned for, use adjusted (for individual-level covariates only) cluster-level outcome summaries and include cluster-level covariates only in TSLS regression. This prevents analysts from having to correct the degrees of freedom themselves.

How to ensure valid inferences?

3. Firstly, use the planned cluster-level ITT analysis and check whether heteroscedasticity exists by comparing the residual variance across trial groups.
   a. Obtain those residuals from a linear regression of the chosen cluster-level outcome summaries approach.
   b. Residuals plot across trial groups can be used, or formal heteroscedasticity tests such as of Breusch-Pagan [94] or White [70].
   c. If there is no evidence of heteroscedasticity across trial groups, no need for weighting.
4. Always use Huber-White standard errors whether heteroscedasticity is present or not, but not when the number of clusters is small [71].
5. Use $t-$distribution rather than normal approximation when the number of clusters is small *i.e.* use small sample degrees of freedom adjustment.

Which weighting strategy to choose?

6. Use minimum-variance weights, except when the number of clusters is small.
7. If the marginal ICC for outcome is large (≥ 0.6 [91]) and number of clusters is small, use "no weight".
8. When clusters are of similar size, minimum-variance, cluster size and "no" weights are equivalent. For simplicity, choose "no weight".

How to interpret cluster-level LATE from TSLS?

9. If adherence is at the cluster level, cluster-level LATE can be interpreted as the population LATE.
10. If non-adherence is at the individual level and clusters are of same size, cluster-level LATE can be interpreted as the population LATE.
11. If non-adherence is at the individual level and clusters are not of same size but LATE is constant across clusters, cluster-level LATE can be interpreted as the

population LATE. To the best of our knowledge, there is no formal procedure to assess the assumption of homogeneous LATE across clusters, which is a strong assumption.

How plausible are LATE identification assumptions?

12. Discuss potential bias introduced by assumptions' violations. Designs such as double blinding enhance the plausibility of LATE identification assumptions [23]. When a threshold of cluster-level proportion of treatment received is used to define adherence to treatment at the cluster level as binary (rather than continuous), analysts should be cautious about departures from the ER assumption. In addition, an incorrect dichotomization of treatment receipt leads to LATE estimates that are too large compared to the average per-unit effect of the continuous treatment received [28].
13. Test the relevance of random treatment assignment as instrument using Staiger & Stock's rule of thumb [106]. The reporting of the instrument's relevance has been advocated [35].
    a. If $F-$statistic$\geq$ 10, random treatment assignment is a relevant instrument and TSLS estimator is consistent.
    b. If $F-$statistic$<$ 10, random treatment assignment is a weak instrument. TSLS estimator may be biased and not appropriate. This is likely to happen when only few clusters are recruited per treatment and non-adherence is at the cluster level. The bias is larger in small samples and very sensitive to even small departures from the ER assumption [76, 104]. In addition, inferences are wrong but robust procedures such as permutation-based inference [130], Anderson-Rubin's structural parameter test [131] and conditional likelihood ratio test [132] have been suggested. Alternatively, seek for estimands other than LATE, for instance average total effects that can be estimated using propensity scores approach (see [133] for binary treatment and [134] for continuous treatment).

Sensitivity analyses to violation of LATE identification assumptions

14. Use for instance TSLS where the original outcome is replaced by a new variable generated as the difference between the original outcome and the fitted interaction between the random treatment assignment and the treatment received. The fitted interaction is an interaction effect used as a sensitivity parameter elicited from external knowledge to enable the point-identification of LATE when relaxing the ER assumption.

Handling of missing data

15. Diaz-Ordaz et al. [42] provided comprehensive guidelines on the assumptions and common techniques that can be used to address missing data when data are clustered. Ignoring clustering in the multiple imputation may inflate the Type I error [135] whereas including cluster's fixed effect in the imputation model may

result in overestimated between-cluster variance in particular in the presence of small cluster sizes and low intracluster correlation coefficient [136].

16. To estimate cluster-level LATE, an option is to first use a multilevel joint modelling multiple imputation [122, 123] assuming a missing-at-random mechanism and impute missing data separately in the control and active groups using the "jomo" package in R [124]. Then, perform cluster-level analysis on each imputed dataset and finally pool the estimates using Rubin's rule [126, 127].
17. The imputation model must include all variables in the analysis model and adding auxiliary variables may enhance the plausibility of the missing-at-random assumption more plausible [42].
18. Evaluate via sensitivity analyses the robustness of the results to departure from the missing-at-random assumption by for instance modifying the multiply-imputed data as follows
    a. Multiply-impute the missing data assuming missing-at-random mechanism.
    b. Modify the imputed data in a. to reflect various plausible scenarios under missing-not-at-random mechanism by multiplying and/or shifting the imputed values by a scalar.
    c. Estimate cluster-level LATE using each modified imputed data and pool the estimates using Rubin's rules [126, 127]. Leurent et al. [137] published a tutorial illustrating the procedure.

After focusing on cluster-level LATE estimation, I compared the performance of the Wald (or conditional Wald) estimator with bootstrapped SEs, TSLS with Huber-White-Rogers SEs, TSLS with Moulton SEs and the Bayesian multilevel mixture modelling, in terms of empirical bias and coverage at the 95% CI, for estimating LATE at individual level. Further illustrations have explored TSLS and Bayesian multilevel mixture model estimating LATE when relaxing the ER assumption.

The Wald estimator with bootstrapped SEs, TSLS with HWR SEs and TSLS with Moulton SEs can be used but the bootstrapped SEs for the Wald estimator tend to result in CIs which are conservative. In the presence of individual-level adherence, when the ICC for outcome is low, TSLS with Moulton's SE is appropriate. However, when the ICC for outcome is large, the Bayesian multilevel model is preferable though it leads to little but negligible bias (below 1%).

As per my findings, I propose some recommendations for trial analysts when individual-

level LATE is of interest and missing data issues have been adequately addressed in a preliminary step, with multiple imputation results obtained using the standard Rubin's rule [126,127]. Those recommendations are summarised in Table 10.2 below.

Table 10.2: Recommendations about the estimation of LATE at individual level

Adherence	Comments	
At CL	If expected ICC for outcome is large	Use either Wald estimator with bootstrapped SEs, TSLS with HWR SEs or TSLS with Moulton's SEs. Covariate adjustment has no effect. Avoid using Bayesian multilevel mixture model if number of clusters is not large.
	If expected ICC for outcome is low	Use either TSLS with HWR SEs, or TSLS with Moulton's SEs. Adjust for CL confounders to improve inference. Avoid using Bayesian multilevel mixture model if number of clusters is not large.
At IL	If expected ICC for outcome is large	Use Bayesian multilevel mixture model. Adjust for CL and/or IL confounders. Use either Uniform or half-Cauchy priors for level-2 standard deviation.
	If expected ICC for outcome is low	Use TSLS with Moulton's SEs with/without covariate adjustment, or Bayesian multilevel mixture model without covariate adjustment. Use either Uniform or half-Cauchy priors for level-2 standard deviation.

CL: cluster level; IL: individual level; TSLS: Two-stage least square; HWR: Huber-White-Rogers; SEs: Standard errors.

Note that when missing data are of concern, the recommendations pertaining to the handling of missing data in Box 10.1 are also applicable to individual-level LATE estimation.

10.2 Strengths and limitations

This research provided the first systematic review of reporting practices of non-adherence with randomised treatment in CRTs and to what extent non-adherence is handled when trial investigators are interested in addressing causal questions. The findings pointed out the need for improving the quality of CRTs reports in terms of causal analyses. The CRTs database used for this review was identified using a rigorous electronic search procedure previously published [42]. This search strategy was calibrated with a previously published one [138], which had been validated with

an ideal set of cluster randomised trials identified from manual examination of a large sample of health journals and was found to have high sensitivity (90.1%). Nevertheless, I may have inherited limitations from the previous strategy search. Some cluster randomised trials may have been missed, as reports may fail to clearly identify the cluster randomisation design in either the title or abstract.

The inclusion criteria were broad, and thus our sample should be representative of the quality of conducting and reporting of CRTs. The included reports were published in 2011, but I do not expect a change in practice for adherence reporting, as the updated CONSORT statement for CRTs [39] was available in pre-print form since 2010 and did not contain any new guidelines with regards to adherence reporting or handling over and above those included in the 2004 version [14]. However, more recent studies focusing on the estimation of causal treatment effects in CRTs [35,98,139] as well as the 2017 addendum to the ICH E9 guidelines [45] encouraging good causal analysis practice have been published and therefore, my review findings may not reflect the current practice.

As with reviews of this nature, our assessments were based only on the information included in the trial reports. It is possible that non-adherence is more common but under-reported, the so-called exclusion paradox [67, 68]. I calculated ranges of non-adherence to reflect this possibility. Another possible limitation is the use of a single reviewer for data extraction. However, single-reviewer extraction was only carried out after a validation phase, where a second reviewer conducted extraction. Agreement between the two extractors was high during validation. Additionally, during full-extraction, whenever there was ambiguity, the second reviewer's opinion was sought and disagreements were resolved by consensus.

I covered the two levels of unit of inference, that is at the cluster level and at the individual level, when addressing non-adherence in CRTs. Cluster-level inference may be preferable for CRTs with a small number of clusters whereas inference at the individual level may be suitable when there is relatively large number of clusters as the between-cluster variance may be adequately estimated [8]. Cluster-level

summary ITT analyses are known to perform well under various settings [8], but the performance of CL-TSLS with alternative weighting while adjusting for cluster-level and/or individual level covariates as well as the Wald estimator with Schochet-Chiang's SEs has not been explored. This research filled this gap by conducting an extensive simulation study where scenarios vary by cluster size and the outcome intracluster correlation coefficient, whether adherence is at the cluster level or at the individual level, and the strength of covariates' effect sizes (small or large effect). Most of the studies investigating cluster size imbalance have only focused on its impact on the statistical power [140–142]. To my knowledge, this research is the first to explore the impact of cluster size imbalance on the coverage of methods for estimating cluster-level LATE.

Selection bias may potentially occur when individuals are recruited after the clusters have been randomised to a treatment group. This does not violate the unconfoundedness assumption, which needs to hold only at the cluster level. It can be argued that random allocation remains a strong instrument, as clusters are more likely to recruit individual participants who find the known cluster allocation desirable/acceptable. However, the validity of the ER is questionable, as in principle, knowledge of the allocation before accepting to participate may have a "direct" effect on the outcome and this should be addressed via sensitivity analyses.

Only IV-based methods allowing for LATE estimation have been explored so far as per the remit of this thesis. A common criticism pertains to the nature of the LATE estimand. The LATE estimand is often criticised because the estimates obtained apply to the "compliers" in the population, and these cannot be observed in practice, thus limiting applicability. However, LATE estimates may be used to provide information about the average causal effect in the entire population [35]. Moreover, the average treatment effect on the compliers is often of interest to patients and medical decision makers, especially when they expect patients to comply with the treatment [143].

The identification assumptions for LATE are often untestable. The assumption

of random treatment assignment is met by design. I simulated that the ER and monotonicity are met. The simulations only considered CRT settings where the random treatment assignment is a strong instrument. In general, the random assignment of treatment is expected to be associated with treatment received in well conducted trials and therefore the findings are relevant for most trial settings in practice. Monotonicity holds in one-sided non-adherence CRT settings like in my simulations. However, I performed sensitivity analyses on real CRT data by relaxing the ER assumption using TSLS and Bayesian multilevel mixture estimation methods. The monotonicity assumption, without which LATE is not point-identifiable, was either met or plausible in the motivating examples. The random treatment assignment and IV relevance assumptions are met in the motivating trials and the ER assumption is plausible. Double blinding design may enhance the plausibility of the identification assumptions.

The simulation study conducted to assess the performance of the Wald (or conditional Wald) estimator, TSLS with Huber-White-Rogers SEs, TSLS with Moulton's SEs and the Bayesian multilevel mixture modelling for estimating LATE at individual level only focused on CRT settings with moderate number of clusters (50 clusters in total). Thus, I did not investigate the small sample properties of those methods. Moreover, the apparent benefit of allowing for level-2 variance heterogeneity across adherence classes rather than across trial groups was only observed from real CRT but has not been explored through simulations. Therefore, my conclusions may not be general but limited to the OPERA trial. The Bayesian multilevel mixture modelling, though very attractive and flexible, may require long iterations to ensure good convergence and therefore may be slow to run.

I did not investigate via simulations situations where the identification assumptions are violated but I only conducted sensitivity analyses to departures from the ER assumption using real CRTs. Missing data are common in trials but my simulations focused on CRTs with complete records. However, missing data were present in one of the motivating example. I used a multilevel joint modelling multiple imputation [122, 123] to handle missing data assuming missing at random and I provided the

code that can be implemented using the "jomo" package in R [124].

I was unable to obtain the results from the Bayesian multilevel mixture model for the TXT4FLUJAB trial where the outcome of interest is binary with a large number of clusters (116 clusters) and number of patients (about 100000) because of convergence and computational issues. Further work is required to optimize the analysis code for such CRT settings.

10.3 Practical implications

This research has compared different methods for estimating LATE at the cluster or individual level in the presence of non-adherence at cluster or individual level. I investigated a wide range of CRT settings combining various numbers of clusters, cluster size levels of ICC for outcome, strengths of cluster-level and individual-level covariates on outcome and adherence to treatment, and LATE effect sizes. The proposed recommendations have covered those CRT settings. The methods explored in this thesis for estimating cluster or individual-level LATE are relatively easy to implement in available statistical software. The provision of analysis codes in Stata and R will facilitate the implementation of these methods by trial investigators and potentially contribute to improving the current suboptimal practice (as per the systematic review conducted as part of this thesis) towards the adequate handling of non-adherence when the assessment of treatment efficacy is of interest. The statistical methods covered in this thesis are applicable to two-level data structures more generally i.e. to observational studies [33, 35]. However, a crucial challenge away from randomised experimental settings is to identify a valid instrumental variable, and thus, each application should carefully argue the validity of the proposed IV.

10.4 Further work

My simulations explored so far CRT settings with continuous outcome when individual-level LATE is of interest whereas I illustrated the methods using trials where the outcomes of interest are continuous and binary. As binary outcomes are often used in CRTs, it would be useful to conduct extensive simulations to assess the performance of TSLS, Wald estimator and Bayesian multilevel mixture model to estimate LATE

at the individual level in the presence of binary outcome. Note that for cluster-level LATE, cluster-level summaries are continuous whether the outcome of interest is binary or continuous. Further simulation study may be needed for assessing the small sample properties of the Wald estimator, TSLS and Bayesian multilevel mixture model when estimating LATE at individual level.

I found little impact of cluster size imbalance when estimating cluster-level LATE. While it is well known that cluster size imbalance reduces the power [140–142], no study to my knowledge has investigated how severe the cluster size imbalance should be to affect the performance of methods in terms of coverage. This may be important for improving the analysis practice as cluster size imbalance is likely in real life.

Moreover, LATE has been the focus of this thesis. It may be relevant to investigate the performance of different methods estimating other estimands such as ATE or ATT when investigators are not interested in LATE. This will offer alternatives for addressing causal questions in CRTs where there is presence of non-adherence.

The lack of simulations when relaxing the ER assumption weakens the in-depth and critical assessment of the performance of sensitivity analyses using TSLS and the Bayesian multilevel mixture modelling under various settings. Extensive simulations assessing the performance of TSLS and the Bayesian multilevel mixture modelling when the ER assumption is violated may be necessary to help understand how severe violations of the ER assumption must be to reverse the results.

My simulation study suggests that when adherence is at cluster level, the Bayesian multilevel mixture modelling performs poorly and slightly less poorly if covariates are adjusted for. This may be because of the relatively small number of clusters in the active group. Note that only clusters in the active group provides information for the prediction of the adherence classes. Having relatively small number of clusters in the active group while increasing the number of parameters to be estimated after adding covariates in the model may reduce the model's predictive accuracy of adherence classes (*i.e.* having to determine to what latent class an individual

belongs but there is little information to do so), which in turn may explain the poor performance of the Bayesian multilevel mixture modelling. Future research is needed to assess the performance of the Bayesian multilevel mixture modelling across a range of number clusters along with different prior's elicitation.

In some CRTs, investigators may be interested in comparing a new treatment to an active comparator instead of a placebo or standard care. When non-adherence occurs in such a setting and there are no other available treatments except the two treatments under investigation, the methods covered in this thesis can be used to estimate LATE. Otherwise, if participants had access to other treatments than those under investigation, this introduces alternative treatments not accounted for at the design stage and weakens the internal validity of the design even if only an intention-to-treat approach is of interest. Methodological developments for addressing causal questions in such CRT settings are needed to help to clarify the assumptions as well as to provide analytical tools to answer such questions. This would be a fruitful avenue for future research.

The data generating process assumes that the treatment effects are homogeneous within principal strata. Note that although the treatment effect was set to a fixed value for all adherent units, the data generating process will reflect settings where the treatment effects moderately vary across units within principal strata as the average treatment effect can be used as a valid measurement in such a population. However, in settings where the treatment effects vary considerably across units within principal strata, future research may help to establish how to allow for treatment effect heterogeneity within the potential outcomes and principal stratification framework and what estimands are identifiable in such settings.

The performance of methods investigated in this thesis could be assessed under various missing data patterns. I do recognise that simulating under the sufficient assumptions must be considered when interpreting our results, and that when violations are suspected, sensitivity analyses must be performed. Simulations exploring settings where the random treatment assignment is only a weak instrument may be

valuable. It may also be worth extending the present research to CRTs' settings where treatment non-adherence may occur at both the cluster and individual levels.

10.5 Conclusion

In this thesis, I have investigated through extensive simulations various IV-based methods for estimating cluster-level LATE and individual-level LATE. I illustrated those methods using two real CRTs and conducted some sensitivity analyses to assess the robustness of the findings to departures from the ER assumption. I have proposed some recommendations pertaining to the estimation of cluster-level LATE and individual-level LATE and for settings where non-adherence occurs at the cluster level or at the individual level. Those recommendations may contribute to improving the quality of analysis when estimating causal treatment effects in the presence of non-adherence in CRTs.

Bibliography

[1] Altman DG. Randomisation. British Medical Journal. 1991;302(6791):1481.

[2] Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? British Medical Journal. 1999;318(7192):1209–1209.

[3] Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. American Journal of Public Health. 2004;94(3):416–422.

[4] Glynn RJ, Brookhart MA, Stedman M, Avorn J, Solomon DH. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. Medical Care. 2007;45(10):S38–S43.

[5] Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. Journal of Clinical Epidemiology. 2015;68(6):603–609.

[6] Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. Bmj. 2003;327(7418):785–789.

[7] Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. BMC medical research methodology. 2005;5(1):10.

[8] Hayes R, Moulton L. Cluster randomised trials. 2009;.

[9] Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London : Arnold; 2000.

[10] Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Annals of internal medicine. 2013;158(3):200–207.

[11] Dodd S, White IR, Williamson P. Nonadherence to treatment protocol in published randomised controlled trials: a review. Trials. 2012;13(1):84.

[12] Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clinical Trials. 2004;1(1):80–90.

[13] Schochet PZ, Chiang HS. Estimation and identification of the complier average causal effect parameter in education RCTs. Journal of Educational and Behavioral Statistics. 2011;36(3):307–345.

[14] Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. British Medical Journal. 2004;328(7441):702–708.

[15] Fisher RA. Statistical methods for research workers. Genesis Publishing Pvt Ltd; 1925.

[16] Pocock SJ, Abdalla M. The hope and the hazards of using compliance data in randomized controlled trials. Statistics in Medicine. 1998;17(3):303–317.

[17] Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391–1398.

[18] White IR. Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research. 2005;14(4):327–347.

[19] Ten Have TR, Normand SLT, Marcus SM, Brown CH, Lavori P, Duan N. Intent-to-Treat vs. Non-Intent-to-Treat Analyses under Treatment Non-Adherence in Mental Health Randomized Trials. Psychiatric Annals. 2008 12;38(12):772–783.

[20] Wertz RT. Intention to treat: once randomized, always analyzed. Clin Aphasiol. 1995;23:57–64.

[21] Heritier SR, Gebski VJ, Keech AC. Inclusion of patients in clinical trial analysis: the intention-to-treat principle. Medical Journal of Australia. 2003;179(8):438–440.

[22] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and statistics. 2004;86(1):4–29.

[23] Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006;17(4):360–372.

[24] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association. 1996;91(434):444–455.

[25] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology. 1974;66(5):688.

[26] Holland PW. Statistics and causal inference. Journal of the American statistical Association. 1986;81(396):945–960.

[27] Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. Econometrica. 1994;62(2):467–475.

[28] Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American statistical Association. 1995;90(430):431–442.

[29] Loeys T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. Biometrics. 2003;59(1):100–105.

[30] Jo B, Asparouhov T, Muthén BO, Ialongo NS, Brown CH. Cluster randomized trials with treatment noncompliance. Psychological Methods. 2008;13(1):1.

[31] Wald A. The fitting of straight lines if both variables are subject to error. The Annals of Mathematical Statistics. 1940;11(3):284–300.

[32] Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999;55(2):463–469.

[33] Conley TG, Hansen CB, Rossi PE. Plausibly exogenous. Review of Economics and Statistics. 2012;94(1):260–272.

[34] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58(1):21–29.

[35] Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Statistics in Medicine. 2014;33(13):2297–2340.

[36] Herrett E, Williamson E, van Staa T, Ranopa M, Free C, Chadborn T, et al. Text messaging reminders for influenza vaccine in primary care: a cluster randomised controlled trial (TXT4FLUJAB). BMJ open. 2016;6(2):e010069.

[37] Underwood M, Lamb S, Eldridge S, Sheehan B, Slowther A. Exercise for depression in care home residents: a randomised controlled trial with cost-effectiveness analysis (OPERA). Health Technology Assessment. 2013;17(18).

[38] Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. Journal of gerontology. 1994;49(2):M85–M94.

[39] Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. British Medical Journal. 2012;.

[40] Cornfield J. Symposium on CHD prevention trials: Design issues in testing life style intervention randomization by group: A formal analysis. American Journal of Epidemiology. 1978;108(2):100–102.

[41] Zhang Z, Peluso MJ, Gross CP, Viscoli CM, Kernan WN. Adherence reporting in randomized controlled trials. Clinical Trials. 2014;11(2):195–204.

[42] Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clinical Trials. 2014;p. 1740774514537136.

[43] Rubin DB. Bayesian inference for causal effects: The role of randomization. The Annals of statistics. 1978;p. 34–58.

[44] Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association. 2005;100(469):322–331.

[45] Agency EM. ICH E9 (R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2017;.

[46] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Medicine.

2010;8(1):1.

[47] Bhutta ZA, Soofi S, Cousens S, Mohammad S, Memon ZA, Ali I, et al. Improvement of perinatal and newborn care in rural Pakistan through community-based strategies: a cluster-randomised effectiveness trial. The Lancet. 2011;377(9763):403–412.

[48] Neuzil KM, Thiem VD, Janmohamed A, Huong VM, Tang Y, Diep NTN, et al. Immunogenicity and reactogenicity of alternative schedules of HPV vaccine in Vietnam: a cluster randomized noninferiority trial. The Journal of the American Medical Association. 2011;305(14):1424–1431.

[49] Tagbor H, Cairns M, Nakwa E, Browne E, Sarkodie B, Counihan H, et al. The clinical impact of combining intermittent preventive treatment with home management of malaria in children aged below 5 years: cluster randomised trial. Tropical Medicine & International Health. 2011;16(3):280–289.

[50] Boorsma M, Frijters DH, Knol DL, Ribbe ME, Nijpels G, van Hout HP. Effects of multidisciplinary integrated care on quality of care in residential care facilities for elderly people: a cluster randomized trial. Canadian Medical Association Journal. 2011;183(11):E724–E732.

[51] Dangour AD, Albala C, Allen E, Grundy E, Walker DG, Aedo C, et al. Effect of a nutrition supplement and physical activity program on pneumonia and walking capacity in Chilean older people: a factorial cluster randomized trial. PLoS Medicine. 2011;8(4):e1001023.

[52] Luoto R, Kinnunen TI, Aittasalo M, Kolu P, Raitanen J, Ojala K, et al. Primary prevention of gestational diabetes mellitus and large-for-gestational-age newborns by lifestyle counseling: a cluster-randomized controlled trial. PLoS Medicine. 2011;8(5):e1001036.

[53] Zamorano J, Erdine S, Pavia A, Kim JH, Al-Khadra A, Westergaard M, et al. Proactive multiple cardiovascular risk factor management compared with usual care in patients with hypertension and additional risk factors: the CRUCIAL trial. Current Medical Research and Opinion. 2011;27(4):821–833.

[54] Acolet D, Allen E, Houston R, Wilkinson AR, Costeloe K, Elbourne D. Improvement in neonatal intensive care unit care: a cluster randomised controlled trial of active dissemination of information. Archives of Disease in Childhood-Fetal and Neonatal Edition. 2011;96(6):F434–F439.

[55] Auger N, Daniel M, Knäuper B, Raynault MF, Pless B. Children and youth perceive smoking messages in an unbranded advertisement from a NIKE marketing campaign: a cluster randomised controlled trial. BMC Pediatrics. 2011;11(1):1.

[56] Beer C, Horner B, Flicker L, Scherer S, Lautenschlager NT, Bretland N, et al. A cluster-randomised trial of staff education to improve the quality of life of people with dementia living in residential care: the DIRECT study. PloS One. 2011;6(11):e28155.

[57] Bickman L, Kelley SD, Breda C, de Andrade AR, Riemer M. Effects of routine feedback to clinicians on mental health outcomes of youths: results of a randomized trial. Psychiatric Services. 2011;.

[58] Cooke LJ, Chambers LC, Añez EV, Croker HA, Boniface D, Yeomans MR, et al. Eating for pleasure or profit the effect of incentives on children's enjoyment of vegetables. Psychological Science. 2011;22(2):190–196.

[59] Cutrer WB, Castro D, Roy KM, Turner TL. Use of an expert concept map as an advance organizer to improve understanding of respiratory failure. Medical Teacher. 2011;33(12):1018–1026.

[60] Estrada CA, Safford MM, Salanitro AH, Houston TK, Curry W, Williams JH, et al. A web-based diabetes intervention for physician: a cluster-randomized effectiveness trial. International Journal for Quality in Health Care. 2011;23(6):682–689.

[61] Smith SM, Paul G, Kelly A, Whitford DL, O'Shea E, O'Dowd T. Peer support for patients with type 2 diabetes: cluster randomised controlled trial. British Medical Journal. 2011;342:d715.

[62] Taveras EM, Gortmaker SL, Hohman KH, Horan CM, Kleinman KP, Mitchell

K, et al. Randomized controlled trial to improve primary care to prevent and manage childhood obesity: the High Five for Kids study. Archives of Pediatrics & Adolescent Medicine. 2011;165(8):714–722.

[63] Zurovac D, Sudoi RK, Akhwale WS, Ndiritu M, Hamer DH, Rowe AK, et al. The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. The Lancet. 2011;378(9793):795–803.

[64] Stiell IG, Nichol G, Leroux BG, Rea TD, Ornato JP, Powell J, et al. Early versus later rhythm analysis in patients with out-of-hospital cardiac arrest. New England Journal of Medicine. 2011;365(9):787–797.

[65] LaBella CR, Huxford MR, Grissom J, Kim KY, Peng J, Christoffel KK. Effect of neuromuscular warm-up on injuries in female soccer and basketball athletes in urban public high schools: cluster randomized controlled trial. Archives of Pediatrics & Adolescent Medicine. 2011;165(11):1033–1040.

[66] Levine DA, Funkhouser EM, Houston TK, Gerald JK, Johnson-Roe N, Allison JJ, et al. Improving Care After Myocardial Infarction Using a 2-Year Internet-Delivered Intervention: The Department of Veterans Affairs Myocardial Infarction–Plus Cluster-Randomized Trial. Archives of Internal Medicine. 2011;171(21):1910–1917.

[67] Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. British Medical Journal. 1996;312(7033):742–744.

[68] Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. The Lancet. 2002;359(9308):781–785.

[69] Ivers N, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. British Medical Journal. 2011;343:d5886.

[70] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980;48(4):817–838.

[71] Imbens GW, Kolesar M. Robust standard errors in small samples: Some practical advice. Review of Economics and Statistics. 2016;98(4):701–712.

[72] Prais SJ, Aitchison J. The grouping of observations in regression analysis. Revue de l'Institut International de Statistique. 1954;p. 1–22.

[73] Lee EW, Dubin N. Estimation and sample size considerations for clustered binary responses. Statistics in Medicine. 1994;13(12):1241–1252.

[74] Kerry SM, Martin Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Statistics in Medicine. 2001;20(3):377–390.

[75] Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Family practice. 2000;17(2):192–196.

[76] Angrist JD, Pischke JS. Mostly harmless econometrics: An empiricist's companion. Princeton university press; 2008.

[77] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer Science & Business Media; 2009.

[78] Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using Stata. STATA press; 2008.

[79] Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971;58(3):545–554.

[80] Cameron AC, Trivedi PK. Microeconometrics: methods and applications. Cambridge university press; 2005.

[81] Frisch R, Waugh FV. Partial time regressions as compared with individual trends. Econometrica: Journal of the Econometric Society. 1933;p. 387–401.

[82] Lovell MC. Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association.

1963;58(304):993–1010.

[83] Filoso V. Regression anatomy, revealed. The Stata Journal. 2013;13(1):92–106.

[84] Kmenta J. Elements of econometrics. 2nd ed. New York : London: Macmillan ; Collier Macmillan; 1986.

[85] Raudenbush SW, Martinez A, Spybrook J. Strategies for improving precision in group-randomized experiments. Educational Evaluation and Policy Analysis. 2007;29(1):5–29.

[86] Konstantopoulos S. The impact of covariates on statistical power in cluster randomized designs: Which level matters more? Multivariate Behavioral Research. 2012;47(3):392–420.

[87] Jakob EM, Marshall SD, Uetz GW. Estimating fitness: a comparison of body condition indices. Oikos. 1996;p. 61–67.

[88] Romano JP, Wolf M. Resurrecting weighted least squares. Journal of Econometrics. 2017;197(1):1–19.

[89] Campbell MJ, Walters SJ. How to design, analyse and report cluster randomised trials in medicine and health related research. John Wiley & Sons; 2014.

[90] Ahn C, Heo M, Zhang S. Sample size calculations for clustered and longitudinal outcomes in clinical research. Chapman and Hall/CRC; 2014.

[91] Ahn C. An evaluation of methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. Biometrical Journal. 1997;39(7):793–807.

[92] Jung SH, Kang SH, Ahn C. Sample size calculations for clustered binary data. Statistics in medicine. 2001;20(13):1971–1982.

[93] Aitkin A. On least squares and linear combination of observations. Proceedings of the Royal Society of Edinburgh. 1935;55:42–48.

[94] Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. Econometrica. 1979;p. 1287–1294.

[95] Kang H, Keele L. Spillover Effects in Cluster Randomized Trials with Noncompliance. arXiv preprint arXiv:180806418. 2018;.

[96] Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. Journal of the American Statistical Association. 2006;101(476):1398–1407.

[97] Frangakis CE, Rubin DB, Zhou XH. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. Biostatistics. 2002;3(2):147–164.

[98] Kang H, Keele L. Estimation Methods for Cluster Randomized Trials with Noncompliance: A Study of A Biometric Smartcard Payment System in India. arXiv preprint arXiv:180503744. 2018;.

[99] Wooldridge JM. Econometric analysis of cross section and panel data. MIT press; 2010.

[100] Vansteelandt S, Didelez V. Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. Scandinavian Journal of Statistics. 2018;.

[101] Little RJ, Long Q, Lin X. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. Biometrics. 2009;65(2):640–649.

[102] Durbin J. Errors in variables. Revue de l'institut International de Statistique. 1954;p. 23–32.

[103] Oehlert GW. A note on the delta method. The American Statistician. 1992;46(1):27–29.

[104] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association. 1995;90(430):443–450.

[105] Kul S, Vanhaecht K, Panella M. Intraclass correlation coefficients for cluster randomized trials in care pathways and usual care: hospital treatment for

heart failure. BMC health services research. 2014;14(1):84.

[106] Staiger DO, Stock JH. Instrumental variables regression with weak instruments. Econometrica. 1997;65(3):557–586.

[107] Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review. 2003;27(1):79–103.

[108] Guittet L, Ravaud P, Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. BMC medical research methodology. 2006;6(1):17.

[109] Muthén L, Muthén B. Mplus. The comprehensive modelling program for applied researchers: user's guide. 2015;5.

[110] Rogers W. Regression standard errors in clustered samples. Stata technical bulletin. 1994;3(13).

[111] Shore-Sheppard L, et al. The precision of instrumental variables estimates with grouped data. vol. 374. Industrial Relations Section, Princeton University Princeton, NJ; 1996.

[112] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986;p. 54–75.

[113] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in medicine. 2000;19(9):1141–1164.

[114] Gelman A, et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis. 2006;1(3):515–534.

[115] Gamerman D, Lopes HF. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall/CRC; 2006.

[116] Brooks S, Gelman A, Jones G, Meng XL. Handbook of markov chain monte carlo. CRC press; 2011.

[117] Plummer M, et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. vol. 124; 2003.

[118] Gelman A, Rubin DB, et al. Inference from iterative simulation using multiple sequences. Statistical science. 1992;7(4):457–472.

[119] Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics. 1998;7(4):434–455.

[120] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977;p. 1–38.

[121] Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). Journal of statistical software. 2012;48(2):1–36.

[122] Carpenter J, Kenward M. Multiple imputation and its application. John Wiley & Sons; 2012.

[123] Quartagno M, Carpenter J. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. Statistics in medicine. 2016;35(17):2938–2954.

[124] Quartagno M, Carpenter J. jomo: A package for multilevel joint modelling multiple imputation. R package version. 2016;p. 2–2.

[125] Yucel RM. Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. Statistical modelling. 2011;11(4):351–370.

[126] Rubin DB. Multiple imputation for nonresponse in surveys. vol. 81. John Wiley & Sons; 2004.

[127] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:b2393.

[128] Donner A. Some aspects of the design and analysis of cluster randomization trials. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1998;47(1):95–113.

[129] Donald SG, Lang K. Inference with difference-in-differences and other panel data. The review of Economics and Statistics. 2007;89(2):221–233.

[130] Imbens GW, Rosenbaum PR. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2005;168(1):109–126.

[131] Anderson TW, Rubin H. Estimation of the parameters of a single equation in a complete system of stochastic equations. The Annals of Mathematical Statistics. 1949;20(1):46–63.

[132] Moreira MJ. A conditional likelihood ratio test for structural models. Econometrica. 2003;71(4):1027–1048.

[133] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[134] Hirano K, Imbens GW. The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives. 2004;226164:73–84.

[135] Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. Biometrical journal. 2008;50(3):329–345.

[136] Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. Biometrical journal. 2011;53(1):57–74.

[137] Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. PharmacoEconomics. 2018;36(8):889–901.

[138] Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. BMC Medical Research Methodology. 2010;10(1):1.

[139] Schochet PZ. Estimators for clustered education RCTs using the Neyman model for causal inference. Journal of Educational and Behavioral Statistics. 2013;38(3):219–238.

[140] Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. American journal of epidemiology. 1981;114(6):906–914.

[141] Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International journal of epidemiology. 2006;35(5):1292–1300.

[142] van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in medicine. 2007;26(13):2589–2603.

[143] Murray EJ, Caniglia EC, Swanson SA, Hernández-Díaz S, Hernán MA. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. Journal of clinical epidemiology. 2018;103:10–21.

Appendices

A.1 Systematic Review Protocol

Title: Reporting non-adherence in cluster randomised trials: A systematic review

Lead author and data extractor:	Schadrac C. Agbla
Supervisor (CRTs selection):	Karla Diaz-Ordaz
Sifting and selection of studies:	Karla Diaz-Ordaz, Claire Coleman, Abie Cohen
Advisors (CRT non-adherence):	Karla Diaz-Ordaz

Background

The aim of this project is to establish the prevalence of non-adherence in cluster randomised trials (CRTs) (how many cluster randomised trials reported non-adherence and if reported, what percentage of non-adherence), how many studies have addressed non-adherence as well as how non-adherence is handled in CRTs reported in 2011 and to establish the occurrence of methods used for adjusting for non-adherence, e.g. intention-to-treat analysis (ITT), per-protocol analysis (PP), as treated (AT), instrument variable, propensity score, principal stratification, complier average causal effect (CACE).

In addition, the frequency of the pattern of non-adherence will be provided (binary, categorical, continuous, accurately measured or not, non-adherence at individual or cluster level).

To this end, we aim to collect information on primary outcome, method of analysis for primary outcome, description of non-adherence. We will extract data on total sample size, number of clusters/individuals randomly allocated in each trial arm and number of clusters/individuals that received intended treatment. The questions addressed by this review are:

- To what extend are CRTs analysis accounting for non-adherence with trial protocol?
- How non-adherence is measured and reported?
- What methods are being used to handle non-adherence?
- Are these methods accounting for clustering appropriately?

Inclusion criteria

We are using a database of CRTs published in 2011, previously used by Diaz-Ordaz et al 2013. Details of the inclusion criteria are given below.

- Studies will be included if they are full report of phase 3 randomised controlled trials where the randomisation is by cluster, as long as there are some outcomes collected at a level below randomisation unit.
- All trial designs will be included.
- English-language publications only.
- All published in peer-reviewed journals in 2011.
- Unpublished trials will be excluded. Pilot and feasibility studies will also be excluded.
- When the same study is reported in more than one paper this will be documented, and data extraction on adjusting for non-adherence on secondary papers will be performed.
- No studies will be excluded on the basis of quality, since the aim is to provide a description of current practice, which will include studies of poor quality.
- Studies using data from CRTs as secondary data source (sub-samples) will be excluded.
- Pilot and feasibility studies will be included.
- Crossover trials will be excluded because of the implications of treatment adherence. In crossover trials, units are randomised to sequence of treatments and not to a single treatment. Units may not adhere to their allocated sequence of treatments but may receive a specific treatment as intended and therefore would correctly contribute to estimating the causal treatment effect of this specific treatment.

Methods

Data extraction

Sifting was previously conducted by DiazOrdaz *et al* [42], from which 132 reports were identified. For those 132 studies, the following data will be recorded if available from the full text of the published paper(s):

- Name of the first author
- Year of publication
- Journal
- Study identification number for paper
- Identification numbers of other papers referring to the same study
- Primary outcome: defined as that specified by the authors or, if not specified, the outcome used in sample size calculations. If no sample size calculation was reported, the first outcome presented in the abstract was considered primary.
- Any harm outcomes?
- What is the type of control comparator (active control? usual practice? placebo?)
- Length of intervention?
- Analysis
  - Method of statistical analysis.
  - Intraclass correlation coefficients for adherence behaviour.
- Number of clusters and individuals randomised (total sample size) and analysed (the number of complete cases).
- Number of individuals and clusters that received randomised treatment
- Numbers of clusters randomised per arm
- Reporting of how non-adherence was addressed and nature of non-adherence and whether any statistical technique used to account for the non-adherence was adjusted for clustering
- Level of adherence (cluster-level adherence? Individual-level adherence? Both cluster and individual-level adherence?)
- Nature of adherence measurement (binary? continuous?)
- ICC for adherence behaviour for individual-level adherence
- Reporting of any reason for addressing non-adherence
- Type of intervention: drug, lifestyle changes (e.g. exercise), educational, surgery, vaccination, etc...
- Proportion of non-adherence at cluster and/or individual level

Analysis

- Information will be produced to show the number and type of studies excluded at various points in the project, and the reasons for exclusion (Prisma flow chart).
- Tables will be produced to show the relationship between various aspects pertaining to statistical quality, i.e. correctly accounting for clustering in analysis and adjusting for for non-adherence. Prevalence of non-adherence will be reported, as well as proportion of studies that addressed non-adherence.

SA will record the trial publication characteristics relating to the quantitative items listed in Table 1 using a piloted, standardised form. In cases of any doubt or ambiguity, the paper will be reviewed and extracted by KDO.

Pilot study

The study will be piloted on about 10% of 2011 identified papers, randomly selected (15 papers). The data extraction form (excel) will be piloted. After piloting, we will refine the data extraction definitions and terms if necessary.

References
Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clinical Trials. 2014;p. 1740774514537136.

Dodd S, White IR, Williamson P. Nonadherence to treatment protocol in published randomised controlled trials: a review. Trials. 2012;13(1):84.

Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC medicine. 2010;8(1):1.

Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. Bmj. 2004;328(7441):702708.

A.2 List of papers included in the systematic review

1. Aboud FE, Akhter S: A cluster-randomized evaluation of a responsive stimulation and feeding intervention in bangladesh. Pediatrics 2011, 127(5):e1191-1197.
2. Aburto NJ, Fulton JE, Safdie M, Duque T, Bonvecchio A, Rivera JA: Effect of a school-based intervention on physical activity: cluster-randomized trial. Medicine and science in sports and exercise 2011, 43(10):1898-1906.
3. Acolet D, Allen E, Houston R, Wilkinson AR, Costeloe K, Elbourne D: Improvement in neonatal intensive care unit care: a cluster randomised controlled trial of active dissemination of information. Archives of disease in childhood Fetal and neonatal edition 2011, 96(6):F434-439.
4. Ali NA, Hammersley J, Hoffmann SP, O'Brien JM, Jr., Phillips GS, Rashkin M, Warren E, Garland A: Continuity of care in intensive care units: a cluster-randomized trial of intensivist staffing. American journal of respiratory and critical care medicine 2011, 184(7):803-808.
5. Al-sheyab N, Gallagher R, Crisp J, Shah S: Peer-led education for adolescents with asthma in Jordan: a cluster-randomized controlled trial. Pediatrics 2012, 129(1):e106-112.
6. Amado Guirado E, Pujol Ribera E, Pacheco Huergo V, Borras JM: Knowledge and adherence to antihypertensive therapy in primary care: results of a randomized trial. Gaceta sanitaria / SESPAS 2011, 25(1):62-67.
7. Atlas SJ, Grant RW, Lester WT, Ashburner JM, Chang Y, Barry MJ, Chueh HC: A cluster-randomized trial of a primary care informatics-based system for breast cancer screening. Journal of general internal medicine 2011, 26(2):154-161.
8. Au DH, Udris EM, Engelberg RA, Diehr PH, Bryson CL, Reinke LF, Curtis JR: A randomized trial to improve communication about end-of-life care among patients with COPD. Chest 2012, 141(3):726-735.
9. Auger N, Daniel M, Knauper B, Raynault MF, Pless B: Children and youth perceive smoking messages in an unbranded advertisement from a NIKE marketing campaign: a cluster randomised controlled trial. BMC pediatrics 2011, 11:26.
10. Ayieko P, Ntoburi S, Wagai J, Opondo C, Opiyo N, Migiro S, Wamae A, Mogoa W, Were F, Wasunna A et al: A multifaceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals: a cluster randomised trial. PLoS medicine 2011, 8(4):e1001018.
11. Bains M, Reynolds PA, McDonald F, Sherriff M: Effectiveness and acceptability of face-to-face, blended and e-learning: a randomised trial of orthodontic undergraduates. European journal of dental education : official journal of the Association for Dental Education in Europe 2011, 15(2):110-117.
12. Beer C, Horner B, Flicker L, Scherer S, Lautenschlager NT, Bretland N, Flett P, Schaper F, Almeida OP: A cluster-randomised trial of staff education to improve the quality of life of people with dementia living in residential care: the DIRECT study. PloS one 2011, 6(11):e28155.

13. Bethge M, Herbold D, Trowitzsch L, Jacobi C: Work status and health-related quality of life following multimodal work hardening: a cluster randomised trial. Journal of back and musculoskeletal rehabilitation 2011, 24(3):161-172.

14. Bhutta ZA, Soofi S, Cousens S, Mohammad S, Memon ZA, Ali I, Feroze A, Raza F, Khan A, Wall S et al: Improvement of perinatal and newborn care in rural Pakistan through community-based strategies: a cluster-randomised effectiveness trial. Lancet 2011, 377(9763):403-412.

15. Bian Y, Xiong H, Zhang L, Tang T, Liu Z, Xu R, Lin H, Xu B: Change in coping strategies following intensive intervention for special-service military personnel as civil emergency responders. Journal of occupational health 2011, 53(1):36-44.

16. Bickman L, Kelley SD, Breda C, de Andrade AR, Riemer M: Effects of routine feedback to clinicians on mental health outcomes of youths: results of a randomized trial. Psychiatric services (Washington, DC) 2011, 62(12):1423-1429.

17. Bin Nisar Y, Hafeez A, Zafar S, Southall DP: Impact of essential surgical skills with an emphasis on emergency maternal, neonatal and child health training on the practice of doctors: a cluster randomised controlled trial in Pakistan. Resuscitation 2011, 82(8):1047-1052.

18. Birkenfeld S, Belfer RG, Chared M, Vilkin A, Barchana M, Lifshitz I, Fruchter D, Aronski D, Balicer R, Niv Y et al: Factors affecting compliance in faecal occult blood testing: a cluster randomized study of the faecal immunochemical test versus the guaiac faecal occult test. Journal of medical screening 2011, 18(3):135-141.

19. Blaya JA, Shin S, Contreras C, Yale G, Suarez C, Asencios L, Kim J, Rodriguez P, Cegielski P, Fraser HS: Full impact of laboratory information system requires direct use by clinical staff: cluster randomized controlled trial. Journal of the American Medical Informatics Association : JAMIA 2011, 18(1):11-16.

20. Bodin MC, Strandberg AK: The Orebro prevention programme revisited: a cluster-randomized effectiveness trial of programme effects on youth drinking. Addiction (Abingdon, England) 2011, 106(12):2134-2143.

21. Boorsma M, Frijters DH, Knol DL, Ribbe ME, Nijpels G, van Hout HP: Effects of multidisciplinary integrated care on quality of care in residential care facilities for elderly people: a cluster randomized trial. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 2011, 183(11):E724-732.

22. Boult C, Reider L, Leff B, Frick KD, Boyd CM, Wolff JL, Frey K, Karm L, Wegener ST, Mroz T et al: The effect of guided care teams on the use of health services: results from a cluster-randomized controlled trial. Archives of internal medicine 2011, 171(5):460-466.

23. Brooker DJ, Argyle E, Scally AJ, Clancy D: The enriched opportunities programme for people with dementia: a cluster-randomised controlled trial in 10 extra care housing schemes. Aging & mental health 2011, 15(8):1008-1017.

24. Brotons C, Soriano N, Moral I, Rodrigo MP, Kloppe P, Rodriguez AI, Gonzalez ML, Arino D, Orozco D, Buitrago F et al: Randomized clinical trial to assess the efficacy of a comprehensive programme of secondary prevention of cardiovascular disease in general practice: the PREseAP study. Revista espanola de cardiologia 2011, 64(1):13-20.

25. Brown K, Dormandy E, Reid E, Gulliford M, Marteau T: Impact on informed choice of offering antenatal sickle cell and thalassaemia screening in primary care: a randomized trial. Journal of medical screening 2011, 18(2):65-75.

26. Cabezas C, Advani M, Puente D, Rodriguez-Blanco T, Martin C: Effectiveness of a stepped primary care smoking cessation intervention: cluster randomized clinical trial (ISTAPS study). Addiction (Abingdon, England) 2011, 106(9):1696-1706.

27. Cameron ID, Kurrle SE, Quine S, Sambrook PN, March L, Chan DK, Lockwood K, Cook B, Schaafsma FF: Improving adherence with the use of hip protectors among older people living in nursing care facilities: a cluster randomized trial. Journal of the American Medical Directors Association 2011, 12(1):50-57.

28. Caria MP, Faggiano F, Bellocco R, Galanti MR: Effects of a school-based prevention program on European adolescents' patterns of alcohol use. The Journal of adolescent health : official publication of the Society for Adolescent Medicine 2011, 48(2):182-188.

29. Chan EC, McFall SL, Byrd TL, Mullen PD, Volk RJ, Ureda J, Calderon-Mora J, Morales P, Valdes A, Kay Bartholomew L: A community-based intervention to promote informed decision making for prostate cancer screening among Hispanic American men changed knowledge and role preferences: a cluster RCT. Patient education and counseling 2011, 84(2):e44-51.

30. Christensen JR, Faber A, Ekner D, Overgaard K, Holtermann A, Sogaard K: Diet, physical exercise and cognitive behavioral training as a combined workplace based intervention to reduce body weight and increase physical capacity in health care workers - a randomized controlled trial. BMC public health 2011, 11:671.

31. Christian P, Labrique AB, Ali H, Richman MJ, Wu L, Rashid M, West KP, Jr.: Maternal vitamin A and beta-carotene supplementation and risk of bacterial vaginosis: a randomized controlled trial in rural Bangladesh. The American journal of clinical nutrition 2011, 94(6):1643-1649.

32. Christie J, Bunting B: The effect of health visitors' postpartum home visit frequency on first-time mothers: cluster randomised trial. International journal of nursing studies 2011, 48(6):689-702.

33. Coles CL, Labrique A, Saha SK, Ali H, Al-Emran H, Rashid M, Christian P, West KP, Jr., Klemm R: Newborn vitamin A supplementation does not affect nasopharyngeal carriage of Streptococcus pneumoniae in Bangladeshi infants at age 3 months. The Journal of nutrition 2011, 141(10):1907-1911.

34. Cooke LJ, Chambers LC, Anez EV, Croker HA, Boniface D, Yeomans MR, Wardle J: Eating for pleasure or profit: the effect of incentives on children's enjoyment of vegetables. Psychological science 2011, 22(2):190-196.

35. Crone MR, Spruijt R, Dijkstra NS, Willemsen MC, Paulussen TG: Does a smoking prevention program in elementary schools prepare children for secondary school? Preventive medicine 2011, 52(1):53-59.

36. Cutrer WB, Castro D, Roy KM, Turner TL: Use of an expert concept map as an advance organizer to improve understanding of respiratory failure. Medical teacher 2011, 33(12):1018-1026.

37. D'Acremont V, Kahama-Maro J, Swai N, Mtasiwa D, Genton B, Lengeler C: Reduction of anti-malarial consumption after rapid diagnostic tests implementation in Dar es Salaam: a before-after and cluster randomized controlled study. Malaria journal 2011, 10:107.

38. Dangour AD, Albala C, Allen E, Grundy E, Walker DG, Aedo C, Sanchez H, Fletcher O, Elbourne D, Uauy R: Effect of a nutrition supplement and physical activity program on pneumonia and walking capacity in Chilean older people: a factorial cluster randomized trial. PLoS medicine 2011, 8(4):e1001023.

39. DiStefano LJ, Blackburn JT, Marshall SW, Guskiewicz KM, Garrett WE, Padua DA: Effects of an age-specific anterior cruciate ligament injury prevention program on lower extremity biomechanics in children. The American journal of sports medicine 2011, 39(5):949-957.

40. Dirks M, Niessen LW, van Wijngaarden JD, Koudstaal PJ, Franke CL, van Oostenbrugge RJ, Huijsman R, Lingsma HF, Minkman MM, Dippel DW: Promoting thrombolysis in acute ischemic stroke. Stroke; a journal of cerebral circulation 2011, 42(5):1325-1330.

41. Driessen MT, Proper KI, Anema JR, Knol DL, Bongers PM, van der Beek AJ: The effectiveness of participatory ergonomics to prevent low-back and neck pain–results of a cluster randomized controlled trial. Scandinavian journal of work, environment & health 2011, 37(5):383-393.

42. Eaton CB, Parker DR, Borkan J, McMurray J, Roberts MB, Lu B, Goldman R, Ahern DK: Translating cholesterol guidelines into primary care practice: a multimodal cluster randomized trial. Annals of family medicine 2011, 9(6):528-537.

43. El-Bassel N, Jemmott JB, 3rd, Landis JR, Pequegnat W, Wingood GM, Wyatt GE, Bellamy SL: Intervention to influence behaviors linked to risk of chronic diseases: a multisite randomized controlled trial with African-American HIV-serodiscordant heterosexual couples. Archives of internal medicine 2011, 171(8):728-736.

44. Epstein JN, Langberg JM, Lichtenstein PK, Kolb R, Altaye M, Simon JO: Use of an Internet portal to improve community-based pediatric ADHD care: a cluster randomized trial. Pediatrics 2011, 128(5):e1201-1208.

45. Esfahanizadeh N: Dental health education programme for 6-year-olds: a cluster randomised controlled trial. European journal of paediatric dentistry : official journal of European Academy of Paediatric Dentistry 2011, 12(3):167-170.

46. Estrada CA, Safford MM, Salanitro AH, Houston TK, Curry W, Williams JH, Ovalle F, Kim Y, Foster P, Allison JJ: A web-based diabetes intervention for physician: a cluster-randomized effectiveness trial. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua 2011, 23(6):682-689.

47. Ettl F, Testori C, Weiser C, Fleischhackl S, Mayer-Stickler M, Herkner H, Schreiber W, Fleischhackl R: Updated teaching techniques improve CPR performance measures: a cluster randomized, controlled trial. Resuscitation 2011, 82(6):730-735.

48. Ezendam NP, Brug J, Oenema A: Evaluation of the Web-based computer-tailored FATaint-PHAT intervention to promote energy balance among adolescents: results from a school cluster randomized trial. Archives of pediatrics & adolescent medicine 2012, 166(3):248-255.

49. Facchin P, Rosa-Rizzotto M, Visona Dalla Pozza L, Turconi AC, Pagliano E, Signorini S, Tornetta L, Trabacca A, Fedrizzi E: Multisite trial comparing the efficacy of constraint-induced movement therapy with that of bimanual intensive training in children with hemiplegic cerebral palsy: postintervention results. American journal of physical medicine & rehabilitation / Association of Academic Physiatrists 2011, 90(7):539-553.

50. Faigenbaum AD, Farrell A, Fabiano M, Radler T, Naclerio F, Ratamess NA, Kang J, Myer GD: Effects of integrative neuromuscular training on fitness performance in children. Pediatric exercise science 2011, 23(4):573-584.

51. Feder G, Davies RA, Baird K, Dunne D, Eldridge S, Griffiths C, Gregory A, Howell A, Johnson M, Ramsay J et al: Identification and Referral to Improve Safety (IRIS) of women experiencing domestic violence with a primary care training and support programme: a cluster randomised controlled trial. Lancet 2011, 378(9805):1788-1795.

52. Fihn SD, Bucher JB, McDonell M, Diehr P, Rumsfeld JS, Doak M, Dougherty C, Gerrity M, Heidenreich P, Larsen G et al: Collaborative care intervention for stable ischemic heart disease. Archives of internal medicine 2011, 171(16):1471-1479.

53. Flather MD, Babalis D, Booth J, Bardaji A, Machecourt J, Opolski G, Ottani F, Bueno H, Banya W, Brady AR et al: Cluster-randomized trial to evaluate the effects of a quality improvement program on management of non-ST-elevation acute coronary syndromes: The European Quality Improvement Programme for Acute Coronary Syndromes (EQUIP-ACS). American heart journal 2011, 162(4):700-707 e701.

54. Foy R, Eccles MP, Hrisos S, Hawthorne G, Steen N, Gibb I, Croal B, Grimshaw J: A cluster randomised trial of educational messages to improve the primary care of diabetes. Implementation science : IS 2011, 6:129.

55. French SA, Gerlach AF, Mitchell NR, Hannan PJ, Welsh EM: Household obesity prevention: Take Action–a group-randomized trial. Obesity (Silver Spring, Md) 2011, 19(10):2082-2088.

56. Galfin JM, Watkins ER, Harlow T: Evaluation of a training programme to teach a guided self-help psychological intervention to hospice staff. International journal of palliative nursing 2011, 17(3):119-124.

57. Goldfeld S, Napiza N, Quach J, Reilly S, Ukoumunne OC, Wake M: Outcomes of a universal shared reading intervention by 2 years of age: the Let's Read trial. Pediatrics 2011, 127(3):445-453.

58. Guldberg TL, Vedsted P, Kristensen JK, Lauritzen T: Improved quality of Type 2 diabetes care following electronic feedback of treatment status to general practitioners: a cluster randomized controlled trial. Diabetic medicine : a journal of the British Diabetic Association 2011, 28(3):325-332.

59. Hanson LC, Carey TS, Caprio AJ, Lee TJ, Ersek M, Garrett J, Jackman A, Gilliam R, Wessell K, Mitchell SL: Improving decision-making for feeding options in advanced dementia: a randomized, controlled trial. Journal of the American Geriatrics Society 2011, 59(11):2009-2016.

60. Hendrie GA, Golley RK: Changing from regular-fat to low-fat dairy foods reduces saturated fat intake but not energy intake in 4-13-y-old children. The American journal of clinical nutrition 2011, 93(5):1117-1127.

61. Hilberink SR, Jacobs JE, Breteler MH, de Vries H, Grol RP: General practice counseling for patients with chronic obstructive pulmonary disease to quit smoking: impact after 1 year of two complex interventions. Patient education and counseling 2011, 83(1):120-124.

62. Holton C, Crockett A, Nelson M, Ryan P, Wood-Baker R, Stocks N, Briggs N, Beilby J: Does spirometry training in general practice improve quality and outcomes of asthma care? International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua 2011, 23(5):545-553.

63. Humphrey LL, Shannon J, Partin MR, O'Malley J, Chen Z, Helfand M: Improving the follow-up of positive hemoccult screening tests: an electronic intervention. Journal of general internal medicine 2011, 26(7):691-697.

64. Huskins WC, Huckabee CM, O'Grady NP, Murray P, Kopetskie H, Zimmer L, Walker ME, Sinkowitz-Cochran RL, Jernigan JA, Samore M et al: Intervention to reduce transmission of resistant bacteria in intensive care. The New England journal of medicine 2011, 364(15):1407-1418.

65. Jackson C, Cheater FM, Harrison W, Peacock R, Bekker H, West R, Leese B: Randomised cluster trial to support informed parental decision-making for the MMR vaccine. BMC public health 2011, 11:475.

66. Jaglal SB, Donescu OS, Bansod V, Laprade J, Thorpe K, Hawker G, Majumdar SR, Meadows L, Cadarette SM, Papaioannou A et al: Impact of a centralized osteoporosis coordinator on post-fracture osteoporosis management: a cluster randomized trial. Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 2012, 23(1):87-95.

67. Jago R, McMurray RG, Drews KL, Moe EL, Murray T, Pham TH, Venditti EM, Volpe SL: HEALTHY intervention: fitness, physical activity, and metabolic syndrome results. Medicine and science in sports and exercise 2011, 43(8):1513-1522.

68. Jemmott JB, 3rd, Jemmott LS, O'Leary A, Ngwane Z, Icard L, Bellamy S, Jones S, Landis JR, Heeren GA, Tyler JC et al: Cognitive-behavioural health-promotion intervention increases fruit and vegetable consumption and physical activity among South African adolescents: a cluster-randomised controlled trial. Psychology & health 2011, 26(2):167-185.

69. Jorgensen MB, Ektor-Andersen J, Sjogaard G, Holtermann A, Sogaard K: A randomised controlled trial among cleaners–effects on strength, balance and kinesiophobia. BMC public health 2011, 11:776.

70. Kaczorowski J, Chambers LW, Dolovich L, Paterson JM, Karwalajtys T, Gierman T, Farrell B, McDonough B, Thabane L, Tu K et al: Improving cardiovascular health at population level: 39 community cluster randomised trial of Cardiovascular Health Awareness Program (CHAP). BMJ (Clinical research ed) 2011, 342:d442.

71. Koczy P, Becker C, Rapp K, Klie T, Beische D, Buchele G, Kleiner A, Guerra V, Rissmann U, Kurrle S et al: Effectiveness of a multifactorial intervention to reduce physical restraints in nursing home residents. Journal of the American Geriatrics Society 2011, 59(2):333-339.

72. Korbkitjaroen M, Vaithayapichet S, Kachintorn K, Jintanothaitavorn D, Wiruchkul N, Thamlikitkul V: Effectiveness of comprehensive implementation of individualized bundling infection control measures for prevention of health care-associated infections in general medical wards. American journal of infection control 2011, 39(6):471-476.

73. LaBella CR, Huxford MR, Grissom J, Kim KY, Peng J, Christoffel KK: Effect of neuromuscular warm-up on injuries in female soccer and basketball athletes in urban public high schools: cluster randomized controlled trial. Archives of pediatrics & adolescent medicine 2011, 165(11):1033-1040.

74. Larisch A, Reuss A, Oertel WH, Eggert K: Does the clinical practice guideline on Parkinson's disease change health outcomes? A cluster randomized controlled trial. Journal of neurology 2011, 258(5):826-834.

75. Law MC, Darrah J, Pollock N, Wilson B, Russell DJ, Walter SD, Rosenbaum P, Galuppi B: Focus on function: a cluster, randomized controlled trial comparing child- versus context-focused intervention for young children with cerebral palsy. Developmental medicine and child neurology 2011, 53(7):621-629.

76. Levine DA, Funkhouser EM, Houston TK, Gerald JK, Johnson-Roe N, Allison JJ, Richman J, Kiefe CI: Improving care after myocardial infarction using a 2-year internet-delivered intervention: the Department of Veterans Affairs myocardial infarction-plus cluster-randomized trial. Archives of internal medicine 2011, 171(21):1910-1917.

77. Li J, Caviness AC, Patel B: Effect of a triage team on length of stay in a pediatric emergency department. Pediatric emergency care 2011, 27(8):687-692.

78. Lim BS, Leung JW, Lee J, Yen D, Beckett L, Tancredi D, Leung FW: Effect of ERCP mechanical simulator (EMS) practice on trainees' ERCP performance in the early learning period: US multicenter randomized controlled trial. The American journal of gastroenterology 2011, 106(2):300-306.

79. Llargues E, Franco R, Recasens A, Nadal A, Vila M, Perez MJ, Manresa JM, Recasens I, Salvador G, Serra J et al: Assessment of a school-based intervention in eating habits and physical activity in school children: the AVall study. Journal of epidemiology and community health 2011, 65(10):896-901.

80. Llor C, Madurell J, Balague-Corbella M, Gomez M, Cots JM: Impact on antibiotic prescription of rapid antigen detection testing in acute pharyngitis in adults: a randomised clinical trial. The British journal of general practice : the journal of the Royal College of General Practitioners 2011, 61(586):e244-251.

81. Lopez-Picazo JJ, Ruiz JC, Sanchez JF, Ariza A, Aguilera B: A randomized trial of the effectiveness and efficiency of interventions to reduce potential drug interactions in primary care. American journal of medical quality : the official journal of the American College of Medical Quality 2011, 26(2):145-153.

82. Luoto R, Kinnunen TI, Aittasalo M, Kolu P, Raitanen J, Ojala K, Mansikkamaki K, Lamberg S, Vasankari T, Komulainen T et al: Primary prevention of gestational diabetes mellitus and large-for-gestational-age newborns by lifestyle counseling: a cluster-randomized controlled trial. PLoS medicine 2011, 8(5):e1001036.

83. MacIntyre CR, Wang Q, Cauchemez S, Seale H, Dwyer DE, Yang P, Shi W, Gao Z, Pang X, Zhang Y et al: A cluster randomized clinical trial comparing fit-tested and non-fit-tested N95 respirators to medical masks to prevent respiratory virus infection in health care workers. Influenza and other respiratory viruses 2011, 5(3):170-179.

84. Magnusson KT, Sigurgeirsson I, Sveinsson T, Johannsson E: Assessment of a two-year school-based physical activity intervention among 7-9-year-old children. The international journal of behavioral nutrition and physical activity 2011, 8:138.

85. McEachan RR, Lawton RJ, Jackson C, Conner M, Meads DM, West RM: Testing a workplace physical activity intervention: a cluster randomized controlled trial. The international journal of behavioral nutrition and physical activity 2011, 8:29.

86. Meyer C, Ulbricht S, Gross B, Kastel L, Wittrien S, Klein G, Skoeries BA, Rumpf HJ, John U: Adoption, reach and effectiveness of computer-based, practitioner delivered and combined smoking interventions in general medical practices: a three-arm cluster randomized trial. Drug and alcohol dependence 2012, 121(1-2):124-132.

87. Middleton S, McElduff P, Ward J, Grimshaw JM, Dale S, D'Este C, Drury P, Griffiths R, Cheung NW, Quinn C et al: Implementation of evidence-based treatment protocols to manage fever, hyperglycaemia, and swallowing dysfunction in acute stroke (QASC): a cluster randomised controlled trial. Lancet 2011, 378(9804):1699-1706.

88. Miller LD, Laye-Gindhu A, Bennett JL, Liu Y, Gold S, March JS, Olson BF, Waechtler VE: An effectiveness study of a culturally enriched school-based CBT anxiety prevention program. Journal of clinical child and adolescent psychology : the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53 2011, 40(4):618-629.

89. Moore Z, Cowman S, Conroy RM: A randomised controlled clinical trial of repositioning, using the 30 degrees tilt, for the prevention of pressure ulcers. Journal of clinical nursing 2011, 20(17-18):2633-2644.

90. Mosnaim GS, Li H, Damitz M, Sharp LK, Li Z, Talati A, Mirza F, Richardson D, Rachelefsky G, Africk J et al: Evaluation of the Fight Asthma Now (FAN) program to improve asthma knowledge in urban youth and teenagers. Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology 2011, 107(4):310-316.

91. Mourad SM, Hermens RP, Liefers J, Akkermans RP, Zielhuis GA, Adang E, Grol RP, Nelen WL, Kremer JA: A multi-faceted strategy to improve the use of national fertility guidelines; a cluster-randomized controlled trial. Human reproduction (Oxford, England) 2011, 26(4):817-826.

92. Neuzil KM, Canh do G, Thiem VD, Janmohamed A, Huong VM, Tang Y, Diep NT, Tsu V, LaMontagne DS: Immunogenicity and reactogenicity of alternative schedules of HPV vaccine in Vietnam: a cluster randomized noninferiority trial. JAMA : the journal of the American Medical Association 2011, 305(14):1424-1431.

93. Ougrin D, Zundel T, Ng A, Banarsee R, Bottle A, Taylor E: Trial of Therapeutic Assessment in London: randomised controlled trial of Therapeutic Assessment versus standard psychosocial assessment in adolescents presenting with self-harm. Archives of disease in childhood 2011, 96(2):148-153.

94. Panunzio MF, Caporizzi R, Antoniciello A, Cela EP, D'Ambrosio P, Ferguson LR, Ruggeri S, Ugolini G, Carella F, Lagravinese D: Training the teachers for improving primary schoolchildren's fruit and vegetables intake: a randomized controlled trial. Annali di igiene : medicina preventiva e di comunita 2011, 23(3):249-260.

95. Peltzer K, Simbayi L, Banyini M, Kekana Q: HIV risk reduction intervention among traditionally circumcised young men in South Africa: a cluster randomized control trial. The Journal of the Association of Nurses in AIDS Care : JANAC 2011, 22(5):397-406.

96. Petersen J, Thorborg K, Nielsen MB, Budtz-Jorgensen E, Holmich P: Preventive effect of eccentric training on acute hamstring injuries in men's soccer: a cluster-randomized controlled trial. The American journal of sports medicine 2011, 39(11):2296-2303.

97. Puder JJ, Marques-Vidal P, Schindler C, Zahner L, Niederer I, Burgi F, Ebenegger V, Nydegger A, Kriemler S: Effect of multidimensional lifestyle intervention on fitness and adiposity in predominantly migrant preschool children (Ballabeina): cluster randomised controlled trial. BMJ (Clinical research ed) 2011, 343:d6195.

98. Regev-Yochay G, Raz M, Dagan R, Roizin H, Morag B, Hetman S, Ringel S, Ben-Israel N, Varon M, Somekh E et al: Reduction in antibiotic use following a cluster randomized controlled multifaceted intervention: the Israeli judicious antibiotic prescription study. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2011, 53(1):33-41.

99. Resnick B, Galik E, Gruber-Baldini A, Zimmerman S: Testing the effect of function-focused care in assisted living. Journal of the American Geriatrics Society 2011, 59(12):2233-2240.

100. Sankaranarayanan R, Ramadas K, Thara S, Muwonge R, Prabhakar J, Augustine P, Venugopal M, Anju G, Mathew BS: Clinical breast examination: preliminary results from a cluster randomized controlled trial in India. Journal of the National Cancer Institute 2011, 103(19):1476-1480.

101. Scales DC, Dainty K, Hales B, Pinto R, Fowler RA, Adhikari NK, Zwarenstein M: A multifaceted intervention for quality improvement in a network of intensive care units: a cluster randomized trial. JAMA : the journal of the American Medical Association 2011, 305(4):363-372.

102. Siega-Riz AM, El Ghormli L, Mobley C, Gillis B, Stadler D, Hartstein J, Volpe SL, Virus A, Bridgman J: The effects of the HEALTHY study intervention on middle school student dietary intakes. The international journal of behavioral nutrition and physical activity 2011, 8:7.

103. Slade GD, Bailie RS, Roberts-Thomson K, Leach AJ, Raye I, Endean C, Simmons B, Morris P: Effect of health promotion and fluoride varnish on dental caries among Australian Aboriginal children: results from a community-randomized controlled trial. Community dentistry and oral epidemiology 2011, 39(1):29-43.

104. Smith SM, Paul G, Kelly A, Whitford DL, O'Shea E, O'Dowd T: Peer support for patients with type 2 diabetes: cluster randomised controlled trial. BMJ (Clinical research ed) 2011, 342:d715.

105. Spijker A, Wollersheim H, Teerenstra S, Graff M, Adang E, Verhey F, Vernooij-Dassen M: Systematic care for caregivers of patients with dementia: a multicenter, cluster-randomized, controlled trial. The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry 2011, 19(6):521-531.

106. Stiell IG, Nichol G, Leroux BG, Rea TD, Ornato JP, Powell J, Christenson J, Callaway CW, Kudenchuk PJ, Aufderheide TP et al: Early versus later rhythm analysis in patients with out-of-hospital cardiac arrest. The New England journal of medicine 2011, 365(9):787-797.

107. Taft AJ, Small R, Hegarty KL, Watson LF, Gold L, Lumley JA: Mothers' AdvocateS In the Community (MOSAIC)- non-professional mentor support to reduce intimate partner violence and depression in mothers: a cluster randomised trial in primary care. BMC public health 2011, 11:178.

108. Tagbor H, Cairns M, Nakwa E, Browne E, Sarkodie B, Counihan H, Meek S, Chandramohan D: The clinical impact of combining intermittent preventive treatment with home management of malaria in children aged below 5 years: cluster randomised trial. Tropical medicine & international health : TM & IH 2011, 16(3):280-289.

109. Taveras EM, Gortmaker SL, Hohman KH, Horan CM, Kleinman KP, Mitchell K, Price S, Prosser LA, Rifas-Shiman SL, Gillman MW: Randomized controlled trial to improve primary care to prevent and manage childhood obesity: the High Five for Kids study. Archives of pediatrics & adolescent medicine 2011, 165(8):714-722.

110. Tomonaga Y, Gutzwiller F, Luscher TF, Riesen WF, Hug M, Diemand A, Schwenkglenks M, Szucs TD: Diagnostic accuracy of point-of-care testing for acute coronary syndromes, heart failure and thromboembolic events in primary care: a cluster-randomised controlled trial. BMC family practice 2011, 12:12.

111. Tylleskar T, Jackson D, Meda N, Engebretsen IM, Chopra M, Diallo AH, Doherty T, Ekstrom EC, Fadnes LT, Goga A et al: Exclusive breastfeeding promotion by peer counsellors in sub-Saharan Africa (PROMISE-EBF): a cluster-randomised trial. Lancet 2011, 378(9789):420-427.

112. van Gaal BG, Schoonhoven L, Mintjes JA, Borm GF, Hulscher ME, Defloor T, Habets H, Voss A, Vloet LC, Koopmans RT et al: Fewer adverse events as a result of the SAFE or SORRY? programme in hospitals and nursing homes. part i: primary outcome of a cluster randomised trial. International journal of nursing studies 2011, 48(9):1040-1048.

113. van Stralen MM, de Vries H, Mudde AN, Bolman C, Lechner L: The long-term efficacy of two computer-tailored physical activity interventions for older adults: main effects and mediators. Health psychology : official journal of the Division of Health Psychology, American Psychological Association 2011, 30(4):442-452.

114. Victor RG, Ravenell JE, Freeman A, Leonard D, Bhat DG, Shafiq M, Knowles P, Storm JS, Adhikari E, Bibbins-Domingo K et al: Effectiveness of a barber-based intervention for improving hypertension control in black men: the BARBER-1 study: a cluster randomized trial. Archives of internal medicine 2011, 171(4):342-350.

115. Vidal J, Borras PA, Ortega FB, Cantallops J, Ponseti X, Palou P: Effects of postural education on daily habits in children. International journal of sports medicine 2011, 32(4):303-308.

116. Vyth EL, Steenhuis IH, Heymans MW, Roodenburg AJ, Brug J, Seidell JC: Influence of placement of a nutrition logo on cafeteria menu items on lunchtime food Choices at Dutch work sites. Journal of the American Dietetic Association 2011, 111(1):131-136.

117. West KP, Jr., Christian P, Labrique AB, Rashid M, Shamim AA, Klemm RD, Massie AB, Mehra S, Schulze KJ, Ali H et al: Effects of vitamin A or beta carotene supplementation on pregnancy-related mortality and infant mortality in rural Bangladesh: a cluster randomized trial. JAMA : the journal of the American Medical Association 2011, 305(19):1986-1995.

118. Yamaoka K, Watanabe M, Hida E, Tango T: Impact of group-based dietary education on the dietary habits of female adolescents: a cluster randomized trial. Public health nutrition 2011, 14(4):702-708.

119. Yarris LM, Fu R, LaMantia J, Linden JA, Gene Hern H, Lefebvre C, Nestler DM, Tupesis J, Kman N: Effect of an educational intervention on faculty and resident satisfaction with real-time feedback in the emergency department. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2011, 18(5):504-512.

120. Zamorano J, Erdine S, Pavia A, Kim JH, Al-Khadra A, Westergaard M, Sutradhar S, Yunis C: Proactive multiple cardiovascular risk factor management compared with usual care in patients with hypertension and additional risk factors: the CRUCIAL trial. Current medical research and opinion 2011, 27(4):821-833.

121. Zebis MK, Andersen LL, Pedersen MT, Mortensen P, Andersen CH, Pedersen MM, Boysen M, Roessler KK, Hannerz H, Mortensen OS et al: Implementation of neck/shoulder exercises for pain relief among industrial workers: a randomized controlled trial. BMC musculoskeletal disorders 2011, 12:205.

122. Zurovac D, Sudoi RK, Akhwale WS, Ndiritu M, Hamer DH, Rowe AK, Snow RW: The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. Lancet 2011, 378(9793):795-803.

123. Zwarenstein M, Fairall LR, Lombard C, Mayers P, Bheekie A, English RG, Lewin S, Bachmann MO, Bateman E: Outreach education for integration of HIV/AIDS care, antiretroviral treatment, and tuberculosis care in primary care clinics in South Africa: PALSA PLUS pragmatic cluster randomised trial. BMJ (Clinical research ed) 2011, 342:d2022.

A.3 Published paper on systematic review

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Schadrac C Agbla
Principal Supervisor	Dr Karla Diaz-Ordaz
Thesis Title	Addressing non-adherence in cluster randomised trials using instrumental variable-based methods

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Clinical Trials		
When was the work published?	April 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I jointly wrote the systematic review protocol, conducted the piloting and validated the data extraction tool with my principal supervisor, Dr Karla Diaz-Ordaz (KDO). I extracted and analysed the data. I jointly drafted the manuscript and approved the final version with KDO.

Student Signature: [redacted] Date: 23/08/2019

Supervisor Signature: [redacted] Date: 23/08/2019

Article

CLINICAL TRIALS

Clinical Trials
1–11

DOI: 10.1177/1740774518761666
journals.sagepub.com/home/ctj
SAGE

Reporting non-adherence in cluster randomised trials: A systematic review

Schadrac C Agbla and Karla DiazOrdaz

Abstract

Background: Treatment non-adherence in randomised trials refers to situations where some participants do not receive their allocated treatment as intended. For cluster randomised trials, where the unit of randomisation is a group of participants, non-adherence may occur at the cluster or individual level. When non-adherence occurs, randomisation no longer guarantees that the relationship between treatment receipt and outcome is unconfounded, and the power to detect the treatment effects in intention-to-treat analysis may be reduced. Thus, recording adherence and estimating the causal treatment effect adequately are of interest for clinical trials.

Objectives: To assess the extent of reporting of non-adherence issues in published cluster trials and to establish which methods are currently being used for addressing non-adherence, if any, and whether clustering is accounted for in these.

Methods: We systematically reviewed 132 cluster trials published in English in 2011 previously identified through a search in PubMed.

Results: One-hundred and twenty three cluster trials were included in this systematic review. Non-adherence was reported in 56 cluster trials. Among these, 19 reported a treatment efficacy estimate: per protocol in 15 and as treated in 4. No study discussed the assumptions made by these methods, their plausibility or the sensitivity of the results to deviations from these assumptions.

Limitations: The year of publication of the cluster trials included in this review (2011) could be considered a limitation of this study; however, no new guidelines regarding the reporting and the handling of non-adherence for cluster trials have been published since. In addition, a single reviewer undertook the data extraction. To mitigate this, a second reviewer conducted a validation of the extraction process on 15 randomly selected reports. Agreement was satisfactory (93%).

Conclusion: Despite the recommendations of the Consolidated Standards of Reporting Trials statement extension to cluster randomised trials, treatment adherence is under-reported. Among the trials providing adherence information, there was substantial variation in how adherence was defined, handled and reported. Researchers should discuss the assumptions required for the results to be interpreted causally and whether these are scientifically plausible in their studies. Sensitivity analyses to study the robustness of the results to departures from these assumptions should be performed.

Keywords

Non-adherence, cluster randomised trials, trial reporting, causal treatment effect

Introduction

Cluster randomised trials, where pre-existing groups of individuals are randomised, have become a common design to test public health and primary care interventions, as often the target of the intervention is a hospital or general practice, or their staff. Increased administrative convenience, reduction of contamination between experimental arms and improved adherence with allocated treatment are often cited among the advantages of adopting this design.[1–3] Nevertheless, treatment adherence may be more challenging in cluster trials because of the hierarchical nature of the design and the delivery of the intervention, where at least two levels at which deviations from protocol called non-adherence can occur, for example, cluster or individual level.[4] The nature of the non-adherence patterns largely depends on

Department of Medical Statistics, London School of Hygiene and Tropical Medicine (LSHTM), London, UK

Corresponding author:
Karla DiazOrdaz, Department of Medical Statistics, London School of Hygiene and Tropical Medicine (LSHTM), Keppel Street, London WC1E 7HT, UK.
Email: karla.diaz-ordaz@lshtm.ac.uk

the nature of the intervention. Some interventions are aimed exclusively at the clusters and thus all individuals within a cluster are exposed to the same treatment. Water fluoridation in a village would be one such example.

In other cluster trials, participants within the same cluster may individually stop adhering to the allocated treatment. An example of this is the study conducted by Sommer et al.,[5] where villages were randomised to 'vitamin A supplements' or not, to be offered to all infants. However, some children whose villages were randomised to 'vitamin A supplements' did not receive the supplements. Finally, non-adherence at both levels is possible for interventions with components targeted at both levels. For example, the OPERA study (exercise for treating depression in care home residents) aimed to evaluate the impact of a 'whole home' exercise intervention on depressive symptoms in older adults living in care homes in England.[6] Clusters were randomly allocated to provide either a depression awareness training session for care home staff (control) or a twice-weekly physiotherapist-led exercise class (active intervention) for 12 months. Some of the exercise classes did not occur due to a shortage of staff (cluster-level non-adherence). In addition, even when the nursing home ran the exercise classes, some individuals recruited to attend these classes did not to do so (individual-level non-adherence). Nursing homes and individuals complied with their assigned treatment during some weeks, but then deviated at a later time, introducing a time-varying non-adherence pattern.

The standard analysis of a trial with departures from allocated treatment is intention-to-treat, which compares outcomes between the groups as randomised, ignoring the actual treatments received. The intention-to-treat estimates the effect of being offered (or allocated to) the treatment and cannot necessarily be interpreted as the causal effect of treatment received. An intention-to-treat analysis with poor adherence may dilute a true treatment effect; with negative outcomes such as side effects, adverse events or mortality, an intention-to-treat estimate which is closer to the null than the true causal effect may make a more toxic or aggressive treatment look less harmful. In addition, non-adherence leads to a loss of power in intention-to-treat analysis.[7]

Where there is an interest in estimating treatment efficacy, as the causal effect of receiving the treatment according to the protocol is called, analytical approaches such as 'per protocol' and 'as treated' are often used.[8] Per protocol restricts the analysis only to participants who received their assigned treatment, whereas as treated analysis compares participants according to their treatment receipt, regardless of their treatment assignment. Both per protocol and as treated may be subject to selection bias, and their validity as causal estimands depends on the assumption that the groups being compared are exchangeable, that is, comparable in terms of their measured and unmeasured covariates. This is a very strong assumption. Since the original comparable groups achieved through randomisation are not preserved, any observed differences in outcomes are not necessarily due to the treatment effect, but potentially also due to differences in covariates.[9] Per protocol also leads to a reduction of statistical power.[7] However, some design features may increase the likelihood of per protocol analysis to be unbiased. One example is when the trial is double-blinded and the treatments have low rates of adverse events, because in this situation, both treatment switching and discontinuations are unlikely to be associated to outcomes, as described by White.[7]

More recently, 'modified' per protocol analyses, that adjust for potential confounders that may lead to selection bias, have been advocated.[10,11] These modified per protocol analyses allow the investigators to adjust for baseline, and post-randomisation variables believed to be sufficient to adjust for the confounding of the association between treatment received and outcome. The assumption of 'no unobserved confounding', required for their validity, is still strong.

There are statistical methods that do not assume 'no unobserved confounding'. Instead, they rely on randomised treatment being an instrumental variable[12–14] and have been proposed in the context of individually randomised trials.[7,15] Extensions to cluster randomised trials exist.[4,16–18] A brief summary of these is given in Box 1. However, methods which are applicable to cluster settings tend to be limited in their usefulness, requiring some programming or the use of specialised software. Previous systematic reviews investigating the reporting and statistical handling of non-adherence in individually randomised controlled trials have been published.[8,26] These have found that adherence to treatment is often under-reported and when reported, sufficient detail on how adherence was defined is often not included. They also found that the majority of the studies used 'unadjusted' per protocol analyses to obtain treatment efficacy estimates.[8] Cluster randomised trials are more complex to run, analyse and report than individually randomised trials, and previous systematic reviews of cluster trials have found that despite the Consolidated Standards of Reporting Trials (CONSORT) extension mentioning explicitly the need to report numbers assigned, on treatment and analysed at the cluster and individual levels, this information is often lacking.[3,27] However, no previous study has focused on the conduct of statistical analyses aiming to estimate causal treatment effects in the presence of treatment non-adherence.

Thus, the aim of this study is to assess the reporting and handling of non-adherence in cluster randomised trials, and in particular, to establish the prevalence of non-adherence and describe the methods used to obtain adherence-adjusted treatment effects. For this, we perform a secondary analysis of data originally extracted for a systematic review investigating the reporting and

Box 1. Causal methods for obtaining adherence-adjusted treatment estimates.

There exists many statistical methods estimating causal treatment effects in randomised trials. They target different estimands, that is, quantities of interest in a defined population, and they also differ in the assumptions required to guarantee identification and unbiased estimation. Interested readers are directed to introductory materials by Bellamy et al.,[15] Baiocchi et al.[18] and Stuart et al.[19] A key idea is that of potential outcomes, that is, the outcome that would have been observed had the randomised allocation been different. Likewise, the potential treatment received is the treatment that individuals/clusters would have received had their randomised allocation been different. Assuming all-or-nothing compliance, the most common assumptions are as follows:

1. *Stable unit* treatment *value assumption*: the potential outcomes of the ith individual are unrelated to the treatment status of all other individuals (known as *no interference*). In addition, we assume *consistency*: for those who actually received treatment-level z, the observed outcome is the potential outcome corresponding to that level of treatment. In cluster trials, stable unit treatment value assumption as above is unlikely to hold. Instead, we may assume that *no interference* holds at the cluster level, that is, the potential outcome of an individual is unrelated to the treatment status of individuals in different clusters, but may depend on those within the same cluster.[4,17]
2. *Ignorability of the treatment assignment*: randomised allocation is independent of unmeasured confounders (conditional on measured covariates) and the potential outcomes.
3. *Instrument relevance*: the random allocation predicts treatment received.
4. *Exclusion restriction*: the random allocation cannot affect the outcomes directly.
5. *Monotonicity*: there are no *defiers*, that is, individuals who receive treatment if and only if they are not randomised to it. Generalisations of these assumptions to cluster trials settings can be found in Schochet.[4]

Here, we concentrate on methods that target the complier average causal effect.

- *Principal stratification.*[14] Under assumptions (1)–(5), each individual may be grouped into a compliance *principal stratum*, which is a latent class, and can be thought of as a baseline covariate:
  1. Never-takers receive no active treatment, regardless of their randomised treatment;
  2. Compliers receive the active treatment if and only if they are randomised to it;
  3. Always-takers, who receive the active treatment, regardless of their randomised treatment.

The complier average causal effect can then be identified from the observed data. Estimation for the principal stratification is based on a mixture of distributions across compliance strata. Extensions to cluster trials are possible, using multilevel mixture models, in either a Bayesian[16] or likelihood approaches.[17]

- *Instrumental variables.* Under assumptions (1)–(5), Angrist et al.[20] showed that the instrumental variable estimand is the intention-to-treat effect in compliers. This is then usually estimated using two-stage least squares. To account for the clustering, it has been recommended to use two-stage least squares using Huber–White variance estimator.[18]
- *Principal scores.* While this method is based on principal stratification, it differs from the previous version, because it does not assume exclusion restriction, but instead assumes *principal ignorability*:[21] the observed covariates are sufficient for identifying principal stratum membership.[22,23] The compliance or principal score is a function solely of pre-randomisation covariates. Similar to the use of propensity scores, this method uses baseline covariates to model principal stratum membership. Once principal scores are obtained, the complier average causal effect is estimated using either matching or weighting, as it is usually done in propensity scores literature. Because it does not assume exclusion restriction, this method is attractive when this assumption is believed to not hold, but the principal ignorability assumption is more plausible.

The choice of causal estimation methods depends on the estimand that investigators are interested in and whether the trial setting supports the plausibility of the underlying causal assumptions. Many of these assumptions are untestable, and often their plausibility is questionable. There are several options to study the sensitivity to departures from these assumptions. For example, when exclusion restriction does not hold, researchers could use the principal scores methods. Alternatively, a Bayesian parametric model can use priors on the non-zero direct effect of randomisation on the outcome for identification on the mixture models used in the principal stratification approach.[24] In the frequentist instrumental variables framework, such modelling is also possible, see Baiocchi et al.[18] for a tutorial on how to perform sensitivity analyses to departures from exclusion restriction and monotonicity assumptions. See also Stuart and Jo[25] for a comparison of the sensitivity to departures from assumptions of principal stratification under exclusion restriction and principal scores under principal ignorability.

adjustment of missing data in cluster trials.[28] We also propose some guidelines for reporting adherence and for conducting adherence-adjusted analysis of cluster trials.

Methods

Search strategy and inclusion criteria

This review uses a database of 132 cluster trials previously identified using a published electronic search strategy.[28] The full electronic search strategy is reported in Box A in the Supplemental File. Reports were eligible for inclusion if they were full reports of cluster randomised controlled trials, published in English in 2011. They were excluded if they were quasi-experimental, self-identified as pilot, feasibility, or preliminary studies; only reported cost-effectiveness or where no data at the individual level were collected. We also excluded cross-over trials, where deviations from randomised treatment may include failure to follow the randomised sequence of treatments, and studies reporting only sub-samples of previously published cluster trials data.

Piloting and validation

Two researchers independently piloted a data extraction form using five randomly selected reports. This

helped to identify extra relevant information to extract and to improve the study protocol. After updating the study protocol and the data extraction form, a random sample of 15 reports was used for validation of the extraction procedures. In case of discrepancy, a final decision was made by consensus and the appropriate information was recorded in the data extraction form. Once the team was satisfied with the extraction procedure, one researcher performed the data extraction in the whole sample. When there was doubt or ambiguity, this was reviewed by the second extractor and a consensus was reached.

Data extraction

Data were extracted on one primary outcome per report, defined as that specified by the authors or, if not specified, the outcome used in sample size calculations. If no primary outcome was specified and no sample size calculation was reported, the first outcome presented in the abstract or manuscript was considered as primary. Information was collected on the type of cluster, the type of primary outcome (binary, continuous and categorical), whether a harm outcome was investigated[29] and the type of intervention given in the control arm (placebo, standard care or active). Information on the level of adherence (cluster level or individual level) was also recorded. Non-adherence was considered to be at the cluster level if the treatment received was different from that assigned for all the participants within clusters, and it was considered to be at the individual level if the treatment received differed from the allocated treatment on an individual basis within the same cluster.

In addition, data on total number of clusters and individuals randomised and analysed were extracted as well as numbers of clusters and individuals receiving the allocated treatment. We defined treatment non-adherence as discrepancy between the allocated course of treatment and the actual treatment received.[8] Descriptions of treatment adherence, including intra-cluster correlation coefficient for treatment adherence,[17] were also recorded, when reported. We also recorded information on adherence-adjusted analyses and whether clustering was accounted for.

We adapted the definitions by Dodd et al.[8] and extracted data about the duration of the intervention. A 'one-off' intervention is defined as that which is received at a single time point, for example, a surgery. A 'short-term' intervention is defined as an intervention implemented at different time points over a short period; for example, five training sessions on the importance of breastfeeding over 1 week. Any other recurrent intervention over an extended period of time was classified as a 'long-term' intervention.

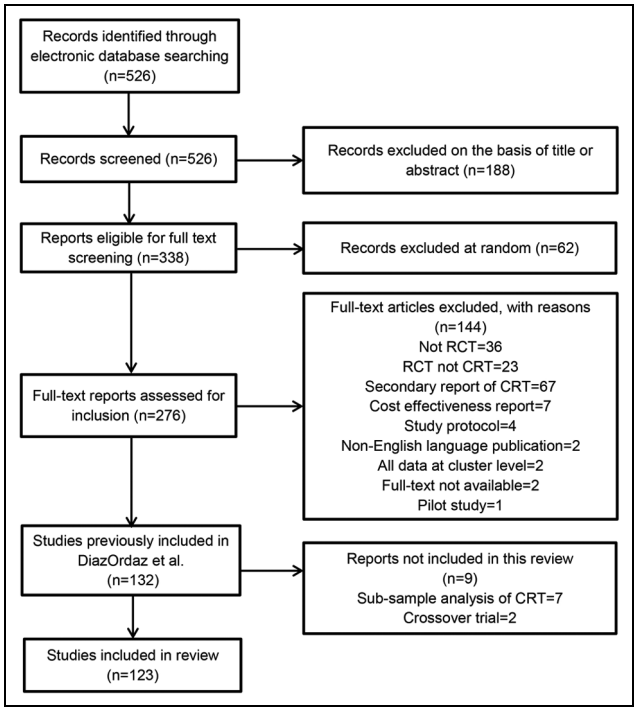


Figure 1. Flow diagram of the identification process for the sample of 123 cluster trials included in this review.

Analysis

Simple analyses were performed to describe the frequency of adherence-reporting and the reported methods used to adjust for non-adherence. The median (and the first and third quartiles) of the number of clusters and individuals randomised, on treatment and analysed, is provided.

For the percentage of non-adherence, we used the author-reported non-adherence when this was reported numerically. If not, we calculated the percentage of non-adherence for each study, from the data provided in the manuscript (the ratio between 'off allocated treatment' participants to the total number randomised).

Results

After excluding seven reports that used only subsamples of cluster trials data and two crossover trials, our final sample included 123 cluster trials. See the flowchart (Figure 1). During the validation phase, the two extractors had an initial agreement of 93%, ultimately achieving consensus by discussion.

Trial characteristics

Trial characteristics are shown in Table 1. Interventions were mainly concerned with changing healthcare practices (63 trials, 51%), educational practices (27 trials,

Table 1. Characteristics of the cluster trials included in this review.

Characteristics	Included trials (123 cluster trials)
Type of intervention, n (%)	
Healthcare practice	63 (51.2)
Lifestyle changes	25 (20.3)
Educational	27 (22.0)
New drug	5 (4.1)
Vaccination/screening	3 (2.4)
Type of control intervention, n (%)	
Standard practice	94 (76.4)
Active control	27 (22.0)
Placebo	2 (1.6)
Primary outcome, n (%)	
Continuous	65 (52.8)
Binary	57 (46.4)
Categorical	1 (0.8)
Investigation of adverse events	12 (9.8)
Number of trial arms, n (%)	
2	106 (86.2)
3–4	17 (13.8)
Level of intervention, n (%)	
Cluster level	65 (52.8)
Individual level	58 (47.2)
Unit of analysis, n (%)	
Clusters	6 (4.9)
Individuals	117 (95.1)
Length of intervention, n (%)	
One-off	5 (4.1)
Short term	35 (28.4)
Long term	83 (67.5)
Clusters randomised per arm, median (first–third quartiles)	12 (7–24)
Range	2–199
Cluster size, median (first–third quartiles)[a]	27 (10–65)
Range[a]	2–14,350
Primary analysis population, n (%)	
Intention-to-treat	119 (96.8)
Per protocol/as treated	4 (3.2)

[a]Based on the average number of individuals per cluster reported in each trials.

22%) or lifestyle (25 trials, 20%). In most trials, standard care was used as the control intervention (96 trials, 76%). The primary outcome was either continuous (65 trials, 53%) or binary (57 trials, 46%), with one exception (multi-category). Adverse events were investigated in 12 trials (10%).

The intervention was implemented exclusively at the cluster level in 65 trials (53%) and at the individual level in 58 trials (47%). Long-term interventions were the most common (83 trials, 68%), followed by short-term interventions (35 trials, 28%). The majority of the studies were two-arm trials (106 trials, 86%). The median (first–third quartiles) number of clusters randomised in each trial arm was 12 (7–24), and the number of clusters per trial arm ranged from 2 to 199. The number of individuals per cluster had a median (first–third quartiles) of 27 (10–65) with a range of 2–14,350.

Intention-to-treat analysis was done as primary analysis in 119 trials (97%), with the remaining 4 trials (3.2%) using per protocol or as treated analysis. Only 6 trials (5%) used cluster-level analysis (primary outcome defined at the cluster level) while the remaining 117 trials used individual-level analysis. Among these, clustering was not accounted for in 12 trials (10%) (see Table 2).

The reporting and handling of non-adherence

Sixty-one reports (50%) included information on adherence: full adherence was reported in five trials while the remaining 56 trials reported some form of non-adherence to the allocated treatment. Table 3 reports the adherence characteristics of these trials. The reporting of adherence was more frequent in interventions of short duration (57%) compared to those of long duration (47%). Forty-four trials (72%) used a binary treatment adherence definition, with only one report justifying the threshold used for this dichotomisation. Only 5 trials (8%) recorded non-adherence as a continuous variable, while the remaining 12 trials (20%) gave no details on the definition of adherence used. Only 11 trials (9%) provided a flowchart with complete information on how many clusters and/or individuals received the assigned treatment. Nine trials reporting non-adherence performed adverse events analysis. Non-adherence at the cluster level was reported in 15 trials (24%), with a further 4 (6%) studies reporting full cluster adherence. Non-adherence at the individual level was reported in 41 trials, representing 71% of the 58 trials reporting treatment non-adherence at the individual level, while 1 trial (2%) reported full adherence. No study reported the use of an intra-cluster correlation coefficient for adherence.

Adherence by allocated groups. *Active group*: 5 studies provided the percentage of cluster-level non-adherence, with a median (first–third quartiles) of 44.8% (33%–50%), with a further 10 reporting cluster non-adherence without further details. At the individual level, 30 (73%) out of 41 studies reported this, with a corresponding median (first–third quartiles) of 15% (9%–24%).

Control group: adherence to the control protocol was less frequently reported; 5 trials stated full adherence, while a further 15 studies reported some form of non-adherence. Cluster-level non-adherence was reported in one trial, while full adherence was reported in a further four trials. At the individual level, 19 trials reported control-treatment non-adherence, with full adherence in 1 study.

Adherence-adjusted analyses. Fifteen trials performed per protocol analyses, with the remaining four studies

Table 2. Analysis methods stratified by unit of analysis.

	Cluster-level analysis	Individual-level analysis
Methods of analysis	6 (100)	117 (100)
Generalised estimating equations	–	27 (23.1)
Mixed effects models	–	55 (47.0)
Repeated measures analysis of variance	–	5 (4.3)
Generalised linear model with sandwich variance	–	16 (13.7)
Chi-square accounting for clustering	–	1 (0.8)
Survival analysis accounting for clustering	–	1 (0.8)
Other methods ignoring clustering[a]	–	12 (10.3)
Weighted regression[b]	1 (16.7)	–
Other methods without weighting[a]	5 (83.3)	–
Methods of analysis when non-adherence was addressed	1 (100)	18 (100)
Generalised estimating equations	–	4 (22.2)
Mixed effects models	–	9 (50.0)
Generalised linear model with sandwich variance	–	4 (22.2)
Poisson regression ignoring clustering[c]	–	1 (5.6)
Unweighted t-test[d]	1 (100)	–

The numbers in brackets are the column percentages.
[a]Generalised linear model, analysis of variance, analysis of covariance, t-test, Mann–Whitney U test and Chi-square test.
[b]Number of events (cluster size) used as weights.[30] The use of weights is applicable when cluster-level summaries analysis is performed while accounting for clustering may be required for individual-level analysis.
[c]t-test with multiple testing adjustment but ignoring clustering was applied to perform a per protocol analysis at individual level.[31]
[d]Per protocol analysis with unweighted t-test comparing rates at cluster level.[32]

carrying out as treated analyses either as primary or secondary analyses. No study reported the complier average causal effect estimate. Among the nine studies with a safety outcome, four trials performed a per protocol analysis,[31,33–35] with a further trial using an as treated analysis.[36] Two studies did not account for clustering in their adherence-adjusted analyses.[31,32] No study reported the assumptions necessary for their adherence-adjusted analyses to be unbiased causal treatment estimates. In any case, none of these studies was double-blinded. We summarise some of the characteristics of these adherence-adjusted analyses in Table 4.

Discussion

This is the first systematic review of reporting practices of non-adherence with randomised treatment in cluster randomised trials. Our findings show that about half of the studies include information on treatment adherence, but details on numbers of clusters and individuals that adhered to the intended treatment are often incomplete. Schulz and colleagues[50,51] found that trials reporting exclusions after treatment initiation (i.e. deviations from protocol) tend to be of higher methodological quality than those that did not report it. This is known as the 'exclusion paradox'. It is, therefore, possible that those studies that did not report on adherence also experienced protocol deviations. On this basis, we estimate that in this study the proportion of trials with non-adherence lies within the range 23%–94% at the cluster level and 71%–98% at the individual level. In addition, we found that studies tended to report more often adherence at the individual level. This potential under-reporting may be due to the complexity of defining adherence in cluster trials and that as cluster trials are often pragmatic in nature, recording adherence to treatment protocol is not often a primary concern.

Among the studies reporting non-adherence, only one-third specified departures from protocol in the control group. This has to be interpreted in light of the fact that in our review, 'usual care' was used as control in approximately three quarters of studies and that defining and measuring adherence to 'usual care' may be difficult or impractical. In general, the nature of the departures from protocol was very poorly reported, and it was not possible to ascertain whether alternative treatments to those in the trial, that is, not originally included in the design of the study, were taken. Knowledge of the alternative regimes followed by those individuals who did not adhere to their allocated treatment is important if we want to judge the impact of such non-adherence on the causal interpretation of an intention-to-treat analysis. If no external treatments are available, then the intention-to-treat estimate will be diluted towards the null, when compared with the true treatment effect. Moreover, the reported non-adherence details (numbers initiating and completing the treatment protocol, period of discontinuation, etc.) were often inadequate for a meaningful interpretation of the study results.

All of the studies reporting adherence-adjusted estimates performed per protocol or as treated analyses, without discussing the plausibility of the necessary assumptions for the results to be interpretable as

Table 3. Reporting of non-adherence by length of intervention, randomised arm and level of adherence.

	One-off	Short term	Long term	Total
Reporting of any non-adherence, n (%)	5 (100)	35 (100)	83 (100)	123 (100)
Non-adherence reported in both active and control groups	–	4 (11.4)	16 (19.3)	20 (16.2)
Non-adherence reported in active group only	2 (40.0)	15 (42.9)	19 (22.9)	36 (29.3)
Non-adherence reported in control group only	–	–	–	–
Full adherence reported	-	1 (2.9)	4 (4.8)	5 (4.1)
Not reported	3 (60.0)	11 (31.4)	36 (43.4)	50 (40.6)
Unclear	–	4 (11.4)	8 (9.6)	12 (9.8)
Trials with adherence at cluster level	2 (100)	21 (100)	42 (100)	65 (100)
Non-adherence reported in both active and control groups	–	–	1 (2.4)	1 (1.5)
Non-adherence reported in active group only	–	9 (42.9)	5 (11.9)	14 (21.5)
Non-adherence reported in control group only	–	–	–	–
Full adherence reported	–	1 (4.7)	3 (7.1)	4 (6.2)
Not reported	2 (100)	9 (42.9)	29 (69.1)	40 (61.6)
Unclear	–	2 (9.5)	4 (9.5)	6 (9.2)
Trials with adherence at individual level	3 (100)	14 (100)	41 (100)	58 (100)
Non-adherence reported in both active and control groups	–	4 (28.6)	15 (36.6)	19 (32.8)
Non-adherence reported in active group only	2 (66.7)	6 (42.9)	14 (34.1)	22 (38.0)
Non-adherence reported in control group only	–	–	–	–
Full adherence reported	–	–	1 (2.4)	1 (1.7)
Not reported	1 (33.3)	2 (14.3)	7 (17.1)	10 (17.2)
Unclear	-	2 (14.3)	4 (9.8)	6 (10.3)
Percentage of non-adherence at cluster level[a]				
Total number of trials reporting non-adherence, n (%)	–	9 (100)	6 (100)	15 (100)
Trials reporting % of non-adherence in active group, n (%)	–	2 (22.2)	3 (50.0)	5 (33.3)
Median % of non-adherence in active group[b]	–	37.4 (30–44.8)	50 (33–80)	44.8 (33–50)
Percentage of non-adherence at individual level				
Total number of trials reporting non-adherence, n (%)	2 (100)	10 (100)	29 (100)	41 (100)
Trials reporting % of non-adherence in active group, n (%)	2 (100)	7 (70.0)	21 (72.4)	30 (73.2)
Median % of non-adherence in active group[b]	16.5 (0.5–32.4)	13.7 (5.3–25)	15 (10–20)	15 (9–24)
Total number of trials reporting non-adherence, n (%)	2 (100)	10 (100)	29 (100)	41 (100)
Trials reporting % of non-adherence in control group, n (%)	–	3 (30.0)	11 (37.9)	14 (34.1)
Median % of non-adherence in control group[b]	–	8.1 (1.7–32)	8.2 (3.4–20)	8.2 (3.4–20)
Total number of trials, n (%)	5 (100)	35 (100)	83 (100)	123 (100)
Flowchart with adherence information	1 (20.0)	4 (11.4)	6 (7.2)	11 (8.9)
Flowchart without adherence information	1 (20.0)	19 (54.3)	65 (78.3)	85 (69.1)
No flow char	3 (60.0)	12 (34.3)	12 (14.5)	27 (22.0)
Adherence type, n (%)[c]	2 (100)	20 (100)	39 (100)	61 (100)
Binary adherence	2 (100)	14 (70.0)	28 (71.8)	44 (72.1)
Continuous adherence	–	2 (10.0)	3 (7.7)	5 (8.2)
Unclear	–	4 (20.0)	8 (20.5)	12 (19.7)
Trials using adherence-adjusted methods, n (%)	1 (100)	4 (100)	14 (100)	19 (100)
Per protocol	1 (100)	4 (100)	10 (71.4)	15 (78.9)
As treated	–	–	4 (28.6)	4 (21.1)

[a]No report provided non-adherence % at cluster level in the control group.
[b]Numbers in brackets are the first and third quartiles.
[c]Total number of trials reporting non-adherence or full adherence.

unbiased causal estimates. These 'unadjusted' analyses rely on the assumption that the association between treatment received and outcome is completely unconfounded.[7,9] Since this assumption is very strong, we would advise adjusting for measured confounders in a 'modified' per protocol analysis (as suggested by Hernán and Robins[11]), thus relying instead on the 'no unobserved confounding' assumption. Moreover, in the context of randomised trials, randomised treatment is very plausibly a valid instrumental variable, and the monotonicity assumption may sometimes hold by design, for example, where the experimental intervention is not available to the controls. These assumptions are sufficient to identify the CACE, allowing the analyst to obtain causal effects in the presence of unobserved confounding. In the absence of complex interactions between compliance classes at the cluster and individual level, statistical methods to estimate CACEs accounting for the clustering in the data could be used (See Box 1). However, no report included in this review performed a complier average causal effect.

Comparison with previous studies

A previous systematic review including 152 cluster trials published between 1997 and 2000 found that non-

Table 4. Details of the adherence-adjusted analyses performed.

Study	Reason	Type	Differences in inference
Per protocol			
Acolet et al.[37]	Exploratory	Binary	PP not shown, stated similar to ITT
Auger et al.[38]	Unclear	Binary	ITT not done
Beer et al.[39]	Unclear	Binary	Evidence of effect with PP, but not with ITT
Bickman et al.[40]	Unclear	Binary	No change
Boorsma et al.[33,a]	Unclear	Binary	Evidence of effect with PP, but not with ITT
Cooke et al.[41]	Unclear	Binary[b]	ITT not done
Cutrer et al.[42]	Unclear	Binary	ITT not done
Dangour et al.[34a]	Exploratory	Binary	No change
Estrada et al.[43]	Unclear	Binary	No change
Luoto et al.[35,a]	Unclear	Binary	No change
Neuzil et al.[31a,c]	Safety	Binary	No change
Smith et al.[44]	Additional analyses	Binary	No change
Tagbor et al.[32,c]	Unclear	Binary[c,d]	Evidence of effect with PP, but not with ITT
Taveras et al.[45]	Unclear	Binary	No change
Zurovac et al.[46]	Unclear	Binary	No change
As treated			
Stiell et al.[47]	Additional analyses	Binary	No change
Zamorano et al.[36,a]	Efficacy	Binary	ITT not done
LaBella et al.[48]	Unclear	Continuous	Evidence of effect with ITT, but not with AT
Levine et al.[49]	Unclear	Continuous	AT not shown

PP: per protocol analysis; ITT: intention to treat analysis.
[a]Carried out a safety outcome analysis.
[b]The threshold chosen to define the binary non-adherence was based on a previous study.
[c]Failed to adjust for clustering in the analysis.
[d]All possible definitions of binary adherence explored ($>$ 1 dose, $>$2 doses and full exposure).

adherence was reported in about 24% of the studies.[52] For individual randomised trials, Dodd et al.,[8] in a review of 100 trials published in 2008 in five leading medical journals, found this to be 98%. In contrast, the review performed by Zhang et al.,[26] which considered individual randomised drug trials published in 2010, found a prevalence of non-adherence reporting of 46%. Both of these results are thus in line with the lower and upper bounds found in our study. These two previous individually randomised trials reviews noted a lack of justification in the threshold used in defining a binary measure of non-adherence.[8,26] In the present review, only one justified this choice.

The median percentage of individual-level non-adherence reported by the cluster trials included here was 13%. Similar median percentages of non-adherence were found in previous reviews of adherence in individually randomised trials, 10%–20% in Dodd et al.[8] and 11.6% in Zhang et al.[26] While the latter reported finding a monotonic trend of adherence with regard to intervention duration,[26] we did not find any such trend. This could be because adherence was not clearly reported in over 40% of both long- and short-term interventions. In fact, in view of the 'exclusion paradox', non-adherence with short-term interventions could be as high as the non-adherence reported in long-term interventions.

Only 3% of the studies included in the present review presented an adherence-adjusted analysis as primary, with the great majority reporting an intention-to-treat approach. Of those studies assessing treatment efficacy, per protocol analysis was the most used. Dodd et al.[8] also found that the majority of studies attempting to adjust for non-adherence in an analysis used per protocol.

Although the extended CONSORT statement for cluster randomised trials[53,54] recommends reporting the numbers of clusters and individuals randomised and receiving their assigned treatment, we found that the reporting of these numbers in the flowchart was low (9%). This is in contrast to the results reported by Dodd et al.,[8] who found that 58% stated the number of participants actually initiating their allocated treatment. A possible explanation may be the lower adherence to CONSORT guidelines among cluster trial reports,[27] as well as the extra complexities of defining, measuring and recording adherence at both levels.

Strengths and limitations

The cluster trials database used for this review was identified using a rigorous electronic search procedure previously published.[28] This search strategy was calibrated with a previously published one,[55] which had been validated with an ideal set of cluster randomised trials identified from manual examination of a large sample of health journals and was found to have high sensitivity (90.1%). Nevertheless, we may have missed

Box 2. Guidelines for analysing and reporting cluster randomised trials with non-adherence to treatment.

1. Report how adherence to treatment is defined and measured. Describe adherence at the cluster and individual level. If dichotomised, justify the choice of threshold made. These choices should be pre-specified in the protocol.[8]
2. Where there is interest in the causal treatment effect, this should be stated clearly in the trial protocol, prior to data collection.
3. Adherence measures should be collected alongside other trial data.
4. Report the number of clusters and individuals that received the intended treatment in each trial arm.[54]
5. Details of the planned causal analyses should be included in the statistical analysis plan, in advance of receiving the data.
6. Efforts should be made in the statistical analysis to reduce any bias introduced by the fact that treatment received may be associated with other variables affecting the outcome.
7. Choose a statistical method that relies on a set of plausible assumptions for the trial at hand and interpret non-adherence adjusted analyses as explanatory.
8. Discuss the assumptions necessary for the chosen analysis method to result in unbiased causal treatment effect estimates and their plausibility in the context of the cluster trial being analysed and reported.
9. In particular, the use of per protocol analysis must be supported by an explanation of why it is reasonable to assume that the group of participants and clusters who did and did not deviate from their allocated treatment are equivalent. If a set of baseline or post-randomisation variables available is believed to be sufficient to adjust for the confounding, a 'modified' per protocol analysis may be valid.[10,11]
10. If clusters or individuals are excluded from analyses, describe whether the fraction excluded is similar between arms and that the included groups were comparable at baseline.[7]
11. Use a method that accounts for clustering adequately. Principal stratification can be used to estimate the complier average causal effect while accounting for clustering; alternatives include multilevel mixture models[17] and Bayesian hierarchical models.[14] Alternatively, instrumental variable methods can use sandwich variance estimation, which is robust to clustering.[18]
12. Sensitivity analyses should be considered when the assumptions necessary for the primary causal analysis are likely to be violated.[16,18]
13. A discussion of potential bias introduced by assumptions' violations in any of the causal analyses should be included in the published report.

some cluster randomised trials, as reports may fail to clearly identify the cluster randomisation design in either title or abstract.

Our inclusion criteria were broad, and thus our sample should be representative of the quality of conducting and reporting of cluster trials. The included reports were published in 2011, but we do not expect a change in practice for adherence reporting, as the updated CONSORT statement for cluster trials[54] was available in pre-print form since 2010 and did not contain any new guidelines with regards to adherence reporting or handling over and above those included in the 2004 version.[53]

As with reviews of this nature, our assessments were based only on the information included in the trial reports. It is possible that non-adherence is more common but under-reported, the so-called exclusion paradox.[50,51] We calculated ranges of non-adherence to reflect this possibility.

Another possible limitation is the use of a single reviewer for data extraction. However, single-reviewer extraction was only carried out after a validation phase, where a second reviewer conducted extraction. Agreement between the two extractors was high during validation. In addition, during full-extraction, whenever there was ambiguity, the second reviewer's opinion was sought and disagreements were resolved by consensus.

Conclusion

Non-adherence with allocated treatment is common in cluster randomised trials, but it is not sufficiently well reported. Our study suggests that cluster-level non-adherence is less common than individual-level non-adherence. However, after taking into consideration possible under-reporting, the overall level of non-adherence in cluster randomised trials may well be comparable with that previously observed in individually randomised trials,[8] weakening the claim that a cluster randomisation design improves treatment-adherence.[1]

A greater effort should be made to improve the quality of reporting of adherence data and analyses. When undertaking causal analyses as part of a cluster randomised trial with non-adherence, researchers should consider carefully the assumptions necessary for their analyses to result in valid inferences and discuss their plausibility in the context of their trial. Sensitivity analyses should be undertaken when departures from these assumptions are suspected. It is also important to remember that in cluster randomised trials, the validity of the results also relies on obtaining an appropriate estimate for the standard errors, for which it is crucial to use a method that correctly models the dependence structure of the data.

Methodologists should make existing causal methods that accommodate the clustering more widely available and easy to implement in commonly used software. To promote their use in practical applications, methodologists should also publish more tutorial papers describing clearly the assumptions needed and detailing the challenges of performing such adherence-adjusted analyses in the context of a good empirical example.

We conclude by making some recommendations for trialists conducting cluster randomised trials (see Box 2).

Previous recommendations for conducting and reporting adherence analyses for individually randomised trials are still relevant.[8] A new framework for the conduct and interpretation of randomised trials in the presence of treatment non-adherence has been recently published, and we encourage the readers to follow these guidelines as much as possible.[56]

Acknowledgements

The authors thank Professor Bianca De Stavola and Dr Clemence Leyrat for their comments on previous drafts of this manuscript. The authors also thank Dr Susanna Dodd for her careful comments and suggestions. K.D.-O. conceived the study. S.C.A. and K.D.-O. wrote the protocol, conducted the piloting and validated the data extraction. S.C.A. extracted and analysed the data. S.C.A. and K.D.-O. jointly drafted the manuscript and approved the final manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This work was supported by the Economic and Social Research Council (grant/award number: ES/J500021/1) and Medical Research Council (grant/award number: MR/L011964/1).

Supplementary material

Electronic search strategy and list of studies included in the review.

ORCID iD

Karla DiazOrdaz https://orcid.org/0000-0003-3155-1561

References

1. Donner A and Klar N. Pitfalls of and controversies in cluster randomization trials. *Am J Public Health* 2004; 94(3): 416–422.
2. Glynn RJ, Brookhart MA, Stedman M, et al. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care* 2007; 45(10): S38–S43.
3. Wright N, Ivers N, Eldridge S, et al. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015; 68(6): 603–609.
4. Schochet PZ and Chiang HS. Estimation and identification of the complier average causal effect parameter in education RCTs. *J Educ Behav Stat* 2011; 36(3): 307–345.
5. Sommer A, Djunaedi E, Loeden A, et al. Impact of Vitamin A supplementation on childhood mortality. A randomised controlled community trial. *Lancet* 1986; 327(8491): 1169–1173.
6. Underwood M, Lamb SE, Eldridge S, et al. Exercise for depression in elderly residents of care homes: a cluster randomised controlled trial. *Lancet* 2013; 382(9886): 41–49.
7. White IR. Uses and limitations of randomization-based efficacy estimators. *Stat Methods Med Res* 2005; 14(4): 327–347.
8. Dodd S, White IR and Williamson P. Nonadherence to treatment protocol in published randomised controlled trials: a review. *Trials* 2012; 13(1): 84.
9. Ten Have TR, Normand SLT, Marcus SM, et al. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatr Ann* 2008; 38(12): 772–783.
10. Murray EJ and Hernn MA. Adherence adjustment in the Coronary Drug Project: a call for better per-protocol effect estimates in randomized trials. *Clin Trials* 2016; 13(4): 372–378.
11. Hernán MA and Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med* 2017; 377(14): 1391–1398.
12. Imbens GW and Angrist JD. Identification and estimation of local average treatment effects. *Econometrica* 1994; 62(2): 467–475.
13. Angrist J and Imbens G. *Identification and estimation of local average treatment effects*. Cambridge, MA: National Bureau of Economic Research, 1995.
14. Frangakis CE and Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; 58(1): 21–29.
15. Bellamy SL, Lin JY and Ten Have TR. An introduction to causal modeling in clinical trials. *Clin Trials* 2007; 4(1): 58–73.
16. Frangakis CE, Rubin DB and Zhou XH. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* 2002; 3(2): 147–164.
17. Jo B, Asparouhov T, Muthén BO, et al. Cluster randomized trials with treatment noncompliance. *Psychol Methods* 2008; 13(1): 1.
18. Baiocchi M, Cheng J and Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014; 33(13): 2297–2340.
19. Stuart EA, Perry DF, Le HN, et al. Estimating intervention effects of prevention programs: accounting for noncompliance. *Prev Sci* 2008; 9(4): 288–298.
20. Angrist JD, Imbens GW and Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; 91(434): 444–455.
21. Ding P and Lu J. Principal stratification analysis using principal scores. *J R Stat Soc B* 2017; 79(3): 757–777.
22. Joffe MM, Ten Have TR and Brensinger C. The compliance score as a regressor in randomized trials. *Biostatistics* 2003; 4(3): 327–340.
23. Jo B and Stuart EA. On the use of propensity scores in principal causal effect estimation. *Stat Med* 2009; 28(23): 2857–2875.
24. Hirano K, Imbens GW, Rubin DB, et al. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000; 1(1): 69–88.
25. Stuart EA and Jo B. Assessing the sensitivity of methods for estimating principal causal effects. *Stat Methods Med Res* 2015; 24(6): 657–674.

26. Zhang Z, Peluso MJ, Gross CP, et al. Adherence reporting in randomized controlled trials. *Clin Trials* 2014; 11(2): 195–204.
27. Ivers N, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ* 2011; 343: d5886.
28. Díaz-Ordaz K, Kenward MG, Cohen A, et al. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014; 11(5): 590–600.
29. Schulz KF, Altman DG and Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med* 2010; 8(1): 18.
30. Bhutta ZA, Soofi S, Cousens S, et al. Improvement of perinatal and newborn care in rural Pakistan through community-based strategies: a cluster-randomised effectiveness trial. *Lancet* 2011; 377(9763): 403–412.
31. Neuzil KM, Thiem VD, Janmohamed A, et al. Immunogenicity and reactogenicity of alternative schedules of HPV vaccine in Vietnam: a cluster randomized noninferiority trial. *JAMA* 2011; 305(14): 1424–1431.
32. Tagbor H, Cairns M, Nakwa E, et al. The clinical impact of combining intermittent preventive treatment with home management of malaria in children aged below 5 years: cluster randomised trial. *Trop Med Int Health* 2011; 16(3): 280–289.
33. Boorsma M, Frijters DH, Knol DL, et al. Effects of multidisciplinary integrated care on quality of care in residential care facilities for elderly people: a cluster randomized trial. *CMAJ* 2011; 183(11): E724–E732.
34. Dangour AD, Albala C, Allen E, et al. Effect of a nutrition supplement and physical activity program on pneumonia and walking capacity in Chilean older people: a factorial cluster randomized trial. *PLoS Med* 2011; 8(4): e1001023.
35. Luoto R, Kinnunen TI, Aittasalo M, et al. Primary prevention of gestational diabetes mellitus and large-for-gestational-age newborns by lifestyle counseling: a cluster-randomized controlled trial. *PLoS Med* 2011; 8(5): e1001036.
36. Zamorano J, Erdine S, Pavia A, et al. Proactive multiple cardiovascular risk factor management compared with usual care in patients with hypertension and additional risk factors: the CRUCIAL trial. *Curr Med Res Opin* 2011; 27(4): 821–833.
37. Acolet D, Allen E, Houston R, et al. Improvement in neonatal intensive care unit care: a cluster randomised controlled trial of active dissemination of information. *Arch Dis Child Fetal Neonatal Ed* 2011; 96(6): F434–F439.
38. Auger N, Daniel M, Knäuper B, et al. Children and youth perceive smoking messages in an unbranded advertisement from a NIKE marketing campaign: a cluster randomised controlled trial. *BMC Pediatr* 2011; 11(1): 26.
39. Beer C, Horner B, Flicker L, et al. A cluster-randomised trial of staff education to improve the quality of life of people with dementia living in residential care: the DIRECT study. *PLoS ONE* 2011; 6(11): e28155.
40. Bickman L, Kelley SD, Breda C, et al. Effects of routine feedback to clinicians on mental health outcomes of youths: results of a randomized trial. *Psychiatr Serv* 2011; 62: 1423–1429.
41. Cooke LJ, Chambers LC, Añez EV, et al. Eating for pleasure or profit the effect of incentives on childrens enjoyment of vegetables. *Psychol Sci* 2011; 22(2): 190–196.
42. Cutrer WB, Castro D, Roy KM, et al. Use of an expert concept map as an advance organizer to improve understanding of respiratory failure. *Med Teach* 2011; 33(12): 1018–1026.
43. Estrada CA, Safford MM, Salanitro AH, et al. A web-based diabetes intervention for physician: a cluster-randomized effectiveness trial. *Int J Qual Health Care* 2011; 23(6): 682–689.
44. Smith SM, Paul G, Kelly A, et al. Peer support for patients with type 2 diabetes: cluster randomised controlled trial. *BMJ* 2011; 342: d715.
45. Taveras EM, Gortmaker SL, Hohman KH, et al. Randomized controlled trial to improve primary care to prevent and manage childhood obesity: the High Five for Kids study. *Arch Pediatr Adolesc Med* 2011; 165(8): 714–722.
46. Zurovac D, Sudoi RK, Akhwale WS, et al. The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. *Lancet* 2011; 378(9793): 795–803.
47. Stiell IG, Nichol G, Leroux BG, et al. Early versus later rhythm analysis in patients with out-of-hospital cardiac arrest. *N Engl J Med* 2011; 365(9): 787–797.
48. LaBella CR, Huxford MR, Grissom J, et al. Effect of neuromuscular warm-up on injuries in female soccer and basketball athletes in urban public high schools: cluster randomized controlled trial. *Arch Pediatr Adolesc Med* 2011; 165(11): 1033–1040.
49. Levine DA, Funkhouser EM, Houston TK, et al. Improving care after myocardial infarction using a 2 year internet-delivered intervention: the Department of Veterans Affairs Myocardial Infarction–Plus Cluster-Randomized Trial. *Arch Intern Med* 2011; 171(21): 1910–1917.
50. Schulz KF, Grimes DA, Altman DG, et al. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* 1996; 312(7033): 742–744.
51. Schulz KF and Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* 2002; 359(9308): 781–785.
52. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004; 1(1): 80–90.
53. Campbell MK, Elbourne DR and Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004; 328(7441): 702–708.
54. Campbell MK, Piaggio G, Elbourne DR, et al. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; 345: e5661.
55. Taljaard M, McGowan J, Grimshaw JM, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Med Res Methodol* 2010; 10(1): 15.
56. Dodd S, White IR and Williamson P. A framework for the design, conduct and interpretation of randomised controlled trials in the presence of treatment changes. *Trials* 2017; 18(1): 498.

A.4 Published paper on CL-LATE estimation

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Schadrac C Agbla
Principal Supervisor	Dr Karla Diaz-Ordaz
Thesis Title	Addressing non-adherence in cluster randomised trials using instrumental variable-based methods

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Statistical Methods in Medical Research		
When was the work published?	May 2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I jointly wrote the simulation plan with Bianca and Karla. I conducted the simulations and analyses of both simulated data and motivating CRTs. I jointly drafted the manuscript and approved the final version with Bianca and Karla.

Student Signature: [redacted] **Date:** 23/08/2019

Supervisor Signature: [redacted] Date: 23/08/2019

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–23

DOI: 10.1177/0962280219849613
journals.sagepub.com/home/smm
SAGE

Estimating cluster-level local average treatment effects in cluster randomised trials with non-adherence

Schadrac C Agbla,[1] Bianca De Stavola[2] and Karla DiazOrdaz[1]

Abstract

Non-adherence to assigned treatment is a common issue in cluster randomised trials. In these settings, the efficacy estimand may also be of interest. Many methodological contributions in recent years have advocated using instrumental variables to identify and estimate the local average treatment effect. However, the clustered nature of randomisation in cluster randomised trials adds to the complexity of such analyses. In this paper, we show that the local average treatment effect can be estimated via two-stage least squares regression using cluster-level summaries of the outcome and treatment received under certain assumptions. We propose the use of baseline variables to adjust the cluster-level summaries before performing two-stage least squares in order to improve efficiency. Implementation needs to account for the reduced sample size, as well as the possible heteroscedasticity, to obtain valid inferences. Simulations are used to assess the performance of two-stage least squares of cluster-level summaries under cluster-level or individual-level non-adherence, with and without weighting and robust standard errors. The impact of adjusting for baseline covariates and of appropriate degrees of freedom correction for inference is also explored. The methods are then illustrated by re-analysing a cluster randomised trial carried out in a specific UK primary care setting. Two-stage least squares estimation using cluster-level summaries provides estimates with small to negligible bias and coverage close to nominal level, provided the appropriate small sample degrees of freedom correction and robust standard errors are used for inference.

Keywords

Cluster randomised trials, non-adherence, local average treatment effect, instrument variable, cluster-level analysis

Introduction

Cluster randomised trials (CRTs), which randomise groups of individuals, are common in public health and primary care. The adoption of this design is often justified given the reduction of 'cross-over contamination' between the experimental arms and improved adherence with allocated treatment.[1–3] Nevertheless, treatment non-adherence is as common in CRTs as it is in individually randomised trials.[4] Dealing with non-adherence is more challenging because there are at least two levels at which deviations from protocol can occur, e.g. cluster or individual level.[5] We say that adherence is at the cluster level if all individuals within a cluster receive the treatment the cluster was randomised to. In contrast, we say that adherence is at the individual level, if the treatment received varies across individuals within the same cluster, so that some individuals received the treatment allocated to their cluster, while others did not.

The standard analysis of randomised clinical trials is intention-to-treat (ITT), which compares average outcomes across randomised groups. However, if the effect of treatment received is confounded, in the sense that there are measured and unmeasured common causes of receiving treatment and experiencing the outcome, the ITT provides the causal effect of being offered, rather than of receiving the treatment. An ITT analysis with poor

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK
[2]Faculty of Population Health Sciences, UCL GOS Institute of Child Health, UK

Corresponding author:
Karla DiazOrdaz, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Email: karla.diaz-ordaz@lshtm.ac.uk

adherence may dilute a true treatment effect.[6] Recently, there has been an increased interest in estimating other estimands alongside the ITT, as highlighted by the International Council for Harmonisation addendum to guideline E9 (Statistical Principles for Clinical Trials). Amongst them, the causal effect in those adhering to treatment has been singled out as being of interest for patients.[7]

In the presence of unmeasured confounding, instrumental variable (IV) methods can estimate consistently the causal effect of an exposure under certain assumptions.[8,9] An IV is a variable which is correlated with the exposure but is not associated with any confounders of the exposure–outcome association, nor is there any pathway by which the IV affects the outcome, other than through the exposure.

Since randomised treatment is usually a valid instrument, IV methods have been proposed to estimate the treatment causal effect in the context of randomised clinical trials affected by non-adherence.[10,11] The population to which an IV estimate applies, however, depends on the assumed behaviour of the instrument.[9] When, as it is often the case, randomised treatment influences treatment received *monotonically*,[9] in the sense that the level of treatment received is greater when randomised to treatment, than when randomised to the control (the precise technical definition will be given shortly), IV methods lead to estimating the causal effect among the adherers, known as the local average treatment effect (LATE) or complier-average causal effect.

This estimand can be estimated via the ratio estimator or the two-stage least squares (TSLS) approach.[12] The latter consists of a 'first stage', which regresses treatment received on randomised treatment, and a 'second stage', which models the outcome on the predicted treatment received. Additional covariates can be included in each stage to control for measured confounding or increase precision. The regression coefficient for the predicted treatment received in the second stage model is a consistent estimator of the LATE, provided that the first stage model is a linear regression, containing all the variables appearing in the second stage.[13,14]

Extensions of this approach for the estimation of LATE in CRTs have been proposed, ranging from a TSLS of individual-level data with variance inflation by the design effect factor[15] to multilevel mixture models that include the latent compliance class membership as a regressor and a random effect for cluster.[16,17] An alternative approach suggested by Schochet[18] constructs Wald-type ratio estimators using cluster-level (CL) summaries for both treatment received and outcome.

In this paper, we focus on TSLS estimation applied to CL outcome summaries. Similar to Schochet,[18] this approach exploits well-known methods from CL analysis, which consist of calculating for each cluster a relevant summary measure of the individual-level outcomes, such as means or proportions, and then analysing these using appropriate statistical methods, such as regression. Because each cluster provides only one data point, the units of analysis can be considered to be independent, but the procedure is inefficient.[19] Estimation by weighted least squares, where the weights are defined either by the cluster size or by the so-called minimum variance weights, can improve the efficiency.[15] Comparing these alternative estimation strategies for the implementation of TSLS estimation using CL data is the focus of this paper. We also demonstrate that using individual-level covariate-adjusted cluster summaries in the (weighted) TSLS regression can increase efficiency.

The rest of the paper proceeds as follows: 'Methodology' section provides an overview of CL analysis methods, defines the estimand of interest and the LATE and introduces the identification assumptions and the different CL TSLS approaches. The finite-sample performance of the methods considered is evaluated using Monte Carlo simulations, presented in 'LATE for CL data' section. In 'Simulation study' section, we illustrate the methods by re-analysing the TXT4FLUJAB trial, a UK-based CRT evaluating the effectiveness and efficacy of text messaging influenza vaccine reminders in increasing vaccine uptake amongst patients with chronic conditions.[20] We conclude with a 'Discussion' section.

Methodology

Consider a two-arm CRT, with n participants, indexed by i, in J clusters, indexed by j, each of size n_j. Let Z_j denotes the binary treatment randomly allocated at the CL with probability 0.50. Let Y_{ij} denotes the continuous or binary outcome, and $D_{ij} \in \{0, 1\}$ be the treatment received by individual i in cluster j. Let W_j and X_{ij} be baseline covariates at CL and individual level, respectively (which can be vectors of variables).

With a slight abuse of notation, we let Y_j denote the CL outcome (mean or proportions), i.e. $Y_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$, hereafter referred to as the unadjusted CL (unCL) outcome. Analogously, let D_j denotes the unCL treatment received, $D_j = \frac{1}{n_j}\sum_{i=1}^{n_j} D_{ij}$.

In the CL adherence settings, D_{ij} is constant within clusters, and therefore D_j is binary. In contrast, when non-adherence is at the individual level, D_j is a continuous measure that varies from 0 to 1, representing the proportion of individuals receiving the active treatment in cluster j.

CL analysis

The unCL analysis uses simple CL summary statistics as the outcomes in subsequent analyses. Let σ^2 denotes the variance of Y_{ij}, which can be decomposed as $\sigma^2 = \sigma_\epsilon^2 + \sigma_\upsilon^2$, where σ_υ^2 is the between-cluster variance and σ_ϵ^2 the within-cluster variance. The intra cluster correlation coefficient (ICC) for Y_{ij} is then $\rho_y = \frac{\sigma_\upsilon^2}{\sigma_\epsilon^2+\sigma_\upsilon^2}$. The variance of Y_j is

$$\begin{aligned} \mathrm{Var}(Y_j) &= \sigma_\upsilon^2 + \frac{\sigma_\epsilon^2}{n_j} = \frac{\sigma_\upsilon^2\{1+(n_j-1)\}+\sigma_\epsilon^2}{n_j} \\ &= \frac{\sigma^2+\sigma_\upsilon^2(n_j-1)}{n_j} = \sigma^2\left\{\frac{1+\rho_y(n_j-1)}{n_j}\right\} \end{aligned} \tag{1}$$

where we have used the fact that $\rho_y = \sigma_\upsilon^2/\sigma^2$ in the last equality.[21]

Since CL outcomes are continuous regardless of whether the original variable was binary, they can be thought to be approximately normally distributed provided n_j is sufficiently large. Thus, a linear regression with CL outcome Y_j as dependent variable and Z_j as the explanatory variable can be fitted to estimate the ITT effects. In the simplest setting without adjustment for other covariates, we have

$$Y_j = \alpha_0 + \alpha_Z Z_j + \eta_j \tag{2}$$

where η_j is a random error term, assumed to be independently and identically distributed (i.i.d.), with mean 0. The ITT is estimated by α_Z.

Efficiency is gained by estimating this model using generalised least squares, with the weights being either the cluster size n_j or the so-called *minimum-variance weights* given by[22]

$$\omega_j = \frac{n_j}{1+\rho_y(n_j-1)}$$

When $\rho_y \approx 0$, minimum-variance weights are approximately equivalent to cluster size weights, while if $\rho_y \approx 1$, minimum-variance weights are approximately 1.[21] These equivalences can have practical implications when the variance of η_j cannot be consistently estimated, for example if the number of clusters is small, so weighting by the cluster size or even no weights are viable alternatives. Where clusters are large, weighting by cluster size is inefficient.[23]

Since the η_j can be heteroscedastic especially when cluster sizes are very imbalanced, the standard errors (SE) should be obtained using a method that takes this into account, such as the Huber–White SE[24] which are consistent when there is heteroscedasticity.[25]

Finally, because each cluster now contributes only one observation, inference should be based on the number of clusters J. Therefore, if p is the number of parameters being estimated, hypothesis tests and confidence intervals (CIs) should be based on appropriate distributions, for example t_{J-p} and not on normal-based approximations. We refer to this as small-sample degrees of freedom (SSDF) correction. Where J is sufficiently large (>40), normal approximations are adequate.

Regression analyses of CL summary outcomes can only adjust for CL covariates directly, as using CL summaries for individual-level regressors is not appropriate.[15] However, where there is interest in adjusting for baseline covariates at the individual level, whether for scientific reasons or to increase statistical efficiency, this can be done through a two-step procedure.[26] First, an individual-level regression analysis of the outcome is performed incorporating all the relevant covariates into the regression model except for the treatment indicator and ignoring clustering, e.g. with only one covariate X_{ij}, we have

$$Y_{ij} = \lambda_0 + \lambda_1 X_{ij} + e_{1_{ij}} \tag{3}$$

In the second step, the sample mean of the fitted residuals for this model $\hat{e}_{1_{ij}}$ is calculated for each cluster j

$$e_j = \frac{1}{n_j}\sum_{j=1}^{n_j} \hat{e}_{1_{ij}}$$

These are then used as CL outcomes in any subsequent analyses. See Appendix 1 for the formulation for binary outcomes.

We refer to these summaries as adCL outcomes. Regression models involving them can also be estimated by generalised least squares, with inference based on normal approximations or Huber–White SEs and/or SSDF corrections, as before. Of note, if CL covariates are used to compute adCL outcomes, the degrees of freedom must be further reduced by the number of CL regressors used to obtain the CL outcome. No such adjustment is necessary for individual-level variables.[26] In this work, we only use adCL outcomes obtained by adjusting for individual-level variables. In the remainder, we denote the CL summary outcomes by Y_j, whether they are unCL or adCL, will be clear from the context.

LATE for CL data

Notation and technical assumptions

Denote by $Y_{ij}(\mathbf{d}_j)$ the *potential outcome* that would manifest if, possibly contrary to fact, the jth cluster to which the individual belongs receives treatment $\mathbf{d_j}$, a vector of length n_j of 0s and 1s, where we are assuming *no interference between clusters*, i.e. the potential outcomes and potential treatment received of individuals in the jth cluster are unrelated to the treatment status of individuals in other clusters.[5] 'No interference between clusters' is a special case of partial interference, where individuals can be partitioned into groups such that interference does not occur between individuals in different groups but may occur between individuals in the same group.[27] This is commonly assumed in clustered randomised trials.[16,17] We also assume *counterfactual consistency*: for $j = 1, \ldots, J$, if $Z_j = z$, then $D_{ij} = D_{ij}(z), \quad Y_{ij} = Y_{ij}(z, D_{ij}(z))$ and $Y_{ij} = Y_{ij}(d)$ for all $i = 1, \ldots, n_j$ and z and d.

The consistency assumption implies that the outcome realised under observation of treatment at level d will equal the potential outcome under a hypothetical intervention to set treatment to value d, regardless of the nature of this hypothetical intervention, in what is called 'treatment-variation irrelevance'.[28,29] More precisely, if we index the different ways of setting the treatment at level d by k_d, the consistency assumption says that $Y_{ij} = Y_{ij}(\mathrm{d}, k_d)$, if $D_{ij} = d_{ij}$, no matter the value of k_d.

Estimand of interest and identification assumptions

Assuming no interference between clusters and consistency allows us to define the estimand of interest, the LATE.[8]

In the setting considered here, where both Z_j and D_{ij} are binary, the vector of potential treatment received under alternative random allocation, $(D_{ij}(0), D_{ij}(1))$ partitions the participants into four different *compliance classes*:[30] $C_{ij} = n$ (never-takers) if $D_{ij}(0) = D_{ij}(1) = 0$; $C_{ij} = a$ (always-takers) if $D_{ij}(0) = D_{ij}(1) = 1$; $C_{ij} = c$ (compliers) if $D_{ij}(z) = z$ for $z \in \{0, 1\}$ and $C_{ij} = d$ (defiers) if $D_{ij}(z) = 1 - z$ for $z \in \{0, 1\}$.

The estimand of interest here is the so-called *population* LATE, defined as

$$\begin{aligned}\beta &= E_j E_i\left[\{Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\} | C_{ij} = c\right] \\ &= \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \{Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\}\{I(D_{ij}(1) = 1, D_{ij}(0) = 0)\}}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} I(D_{ij}(1) = 1, D_{ij}(0) = 0)}\end{aligned} \tag{4}$$

This is said to be a 'local' causal effect as it is conditional on the stratum of complier individuals.

Following Schochet and Chiang,[5,31] we write the cluster version of the corresponding identification assumptions[9] as follows:

A1. CL unconfoundedness: $Z_j \perp\!\!\!\perp D_{ij}(z), Y_{ij}(z, D_{ij}(z)), \quad z \in \{0, 1\}$.

This is also known as cluster randomisation assumption and is often stated in terms of the cluster randomisation to treatment being independent of measured and unmeasured confounders of the relationship between the treatment-received and the outcome. In the context of cluster randomised trials, we know this holds by design.

A2. Exclusion restriction at the individual level: Conditional on the treatment received $D_{ij} = d$, the treatment assignment Z_j has no effect on the outcome. In terms of potential outcomes, we have

$$Y_{ij}(1, d) = Y_{ij}(0, d), \quad \forall d \in \{0, 1\}$$

A3. Instrument relevance: Also referred to as first stage assumption:

Z_j is causally associated with treatment received D_{ij}, i.e. $Z_j \not\perp\!\!\!\perp D_{ij}$.

We remark that in the standard TSLS literature, a weaker version of the assumption A3 is made instead, namely that the instrument Z and treatment received D are only associated, but not necessarily causally. Denote this assumption by A3'. In the causal inference literature, an 'associational' instrument is known as proxy or surrogate instrument. Now, if Z and D are associated but not causally associated, there exists a common cause V, which is the causal instrument, and may be unobserved. In order to define, identify and interpret the LATE causally, under A1, A2 and A3', we further require that Z is conditionally independent from D and Y given V and that V is binary. We refer the interested reader to Hernàn and Robins[32] for further details.

For point identification of local treatment effects, A4 monotonicity of the treatment mechanism is often assumed: $D_{ij}(1) \geq D_{ij}(0)$. In the case of a causal binary instrument, as randomised treatment, the monotonicity asumption implies that there are no individuals who would have received the active treatment when randomised to control ($Z=0$) and not received it when randomised to it. This assumption is often referred informally to as 'there are no defiers'.[8] We remark that the definition and interpretation of the monotonicity assumption is also more complex in non-causal instrument settings.[33] The monotonicity assumption is often justified by design, when the active treatment is not available to those in the control group. Where this is not the case, the investigators have to argue carefully why monotonicity is still plausible.

We also remark that in the case of CL randomisation, but with individual-level non-adherence, we need to assume that monotonicity holds at the individual level.[31] For the CL non-adherence setting, where D_{ij} does not vary within clusters, then this becomes monotonicity at the CL, i.e. $D_j(1)=1$ and $D_j(0)=0$.

An extra assumption necessary when using adCL is that the model used to derive them is correctly specified.

Cluster and individual-level non-adherence

The population-level LATE estimand β can be thought of as a weighted average of the cluster-specific LATE β_j for each cluster j, namely

$$\beta = \sum_{j=1}^{J} \psi_j \beta_j \tag{5}$$

where

$$\begin{aligned} \beta_j &= E_i\big[Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)) | C_{ij} = c\big] \\ &= \frac{1}{n_{c,j}} \sum_{i=1}^{n_j} \Big[\Big\{Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\Big\}\Big\{I(D_{ij}(1) = 1, D_{ij}(0) = 0)\Big\}\Big] \end{aligned} \tag{6}$$

with $n_{c,j}$ is the number of individual-level compliers in each cluster j, assumed here to be >0 for all clusters. The weights corresponding to each β_j are $\psi_j = \frac{n_{c,j}}{\sum_{j=1}^{J} n_{c,j}}$, i.e. the number of cluster-specific compliers divided by the total number of compliers.

This result is useful when interpreting the estimates obtained using CL summaries. We first note that the Wald ratio estimand applied to CL summaries, β_{CL}, does not always correspond to the population LATE β. The former can be expressed as[5]

$$\beta_{CL} = \frac{E[Y_j | Z_j = 1] - E[Y_j | Z_j = 0]}{E[D_j | Z_j = 1] - E[D_j | Z_j = 0]} \tag{7}$$

In the case where treatment received is at the CL (i.e. CL adherence), this CL Wald estimand indeed can be interpreted as the population LATE.

In the case where non-adherence varies at the individual level, it can be shown that $\beta_{CL} = \sum_{j=1}^{J} \psi_{CL,j} \beta_j$ where the CL-weights are $\psi_{CL,j} = \frac{n_{c,j}/n_j}{\sum_j n_{c,j}/n_j}$, i.e. the normalised proportion of individual compliers in each cluster.[31] So, CL–LATE β_{CL} identifies the population LATE, equation (5), only if (i) the cluster sizes n_j are identical for all j or (ii) the cluster-specific LATEs β_j are the same across all clusters, i.e. $\beta_j = \beta$, for all $j \in \{1, \ldots, J\}$.

In the remainder, with individual-level non-adherence, we assume that every cluster has the same cluster-specific LATE, but allow for the cluster sizes to vary. If this is not the case, the CL–LATE β_{CL} identifies a weighted average of the heterogeneous cluster-specific LATE because of clusters with the same proportions of compliers are weighted the same, without accounting for the cluster size.

TSLS for CL data

The conditional expectations appearing in the Wald estimand (equation (7)) can be estimated via standard TSLS regression of the CL summaries (referred to as CL–TSLS). CL–TSLS is most easily explained for settings without weights or covariate adjustment. The first stage fits a regression to CL treatment received D_j on treatment assigned Z_j. Then, in a second stage, a regression for the CL outcome on the predicted treatment received is fitted. This can use either unCL summaries or adCL summaries if there are baseline individual-level variables predictive of the outcome, as this can help gain efficiency. Crucially, both first and second stages must be linear models for the TSLS estimator to be guaranteed to be consistent.[14,34] We have:

$$\begin{aligned} D_j &= \gamma_0 + \gamma_Z Z_j + \omega_{1j} \\ Y_j &= \beta_0 + \beta_{IV}\widehat{D}_j + \omega_{2j} \end{aligned} \tag{8}$$

where ω_{1j} and ω_{2j} are assumed i.i.d. with mean zero and constant variance and such that $\omega_{1j} \perp\!\!\!\perp \omega_{2j}$. The estimate of CL–LATE is then given by $\hat{\beta}_{IV}$.

The asymptotic variance of this estimator is given by $\frac{\widehat{Cov(Y, E[D|Z])}}{\widehat{Cov(D, E[D|Z])}}$, where we assume that $Cov(D, E[D|Z]) \neq 0$. Assuming that the residuals from the second stage $\epsilon = Y - E[Y] - \beta_{IV}D - E[D|Z])$ are such that $E[\epsilon^2|Z = z] = \sigma_Y^2$, the asymptotic variance of the IV estimator simplifies to $\sigma^2(D^\top P_Z D)^{-1}$, where P_Z is the projection matrix $P_Z = Z(Z^\top Z)^{-1}Z^\top Z$. See Imbens and Angrist[8] for further details. These variance estimators are used in most commonly used software implementations of TSLS.

CL covariates can be included in the regressions to increase precision. For example, with one CL covariate W_j, we have

$$\begin{aligned} D_j &= \gamma_0 + \gamma_Z Z_j + \gamma_W W_j + \upsilon_{1j} \\ Y_j &= \beta_0 + \beta_{IV}\hat{D}_j + \beta_W W_j + \upsilon_{2j} \end{aligned} \tag{9}$$

where the error terms are as before.

As with the CL estimation of ITT, where the number of clusters is small, the CIs are constructed using a t distribution with degrees of freedom equal to $J - p$, where p is the number of parameters estimated by the second stage (i.e. the SSDF correction). As before, minimum-variance or cluster size weights can be used to increase efficiency. Finally, the error terms in the CL–TSLS are assumed to be homoscedastic. Where this is not a sensible assumption, Huber–White SEs should be used.[35]

Simulation study

We now perform a simulation study comparing the finite sample performance of TSLS estimation applied to CL data. We simulate CRT individual-level data assuming that the control group does not have access to the active intervention, referred to as one-way non-compliance, at either CL or individual level. In this setting, there are only two compliance classes: compliers and never takers. With a fixed expected total sample size equal to 1000, we vary the number of clusters J and the average cluster size n_j. The marginal ICC of Y also takes two values. The effect of individual and CL variables on the outcome and the treatment received also varies, so that the strength of the confounding is either low or high, while the value of the true LATE also has two levels. Table 1 summarises the factorial design and the values taken by the different levels.

More specifically, we simulate cluster randomised treatment $Z_j \sim \text{Bern}(0.5)$ and two independent baseline covariates, a CL covariate $W_j \sim N(0, \sigma_W^2)$ and individual-level covariate $X_{ij} \sim N(0, \sigma_X^2)$ with a moderate ICC $\rho_X = 0.05$ and with variance $\sigma_W^2 = \sigma_X^2 = 0.08$.

We then generate a binary adherence class indicator variable C_{ij}, which is considered as latent. Let $C_{ij} = 1$ for the compliers, 0 otherwise. For settings where adherence is at the CL, this is constant within clusters, under the following model

$$\begin{aligned} C_{ij} = C_j \sim \text{Bern}(\pi_j) \quad \text{with} \quad \pi_j = P(C_j = 1) \\ \text{logit}(\pi_j) = \lambda_0 + \lambda_W W_j, \end{aligned}$$

with $\lambda_W = 0.05$ equivalent to an odds ratio (OR) ≈ 1.05 per unit increase in W (denoted 'small effect') and $\lambda_W = 0.7$ equivalent to OR ≈ 2 ('large effect').

For settings with individual-level adherence, the data generating model is

$$\begin{aligned} C_{ij} &\sim \text{Bern}(\pi_{ij}) \quad \text{with} \quad \pi_{ij} = \pi = P(C_{ij} = 1) \\ \text{logit}(\pi_{ij}) &= \lambda_0 + \lambda_W W_j + \lambda_X X_{ij} + \zeta_j \\ \zeta_j &\sim N(0, \sigma_\zeta^2) \end{aligned}$$

with $\sigma_\zeta^2 = \pi^2/3$, so that the ICC for compliance is $\rho_C = \sigma_\zeta^2/(\sigma_\zeta^2 + \pi^2/3) = 0.50$.

We derive treatment received at the individual level as

$$D_{ij} = C_{ij} Z_j$$

so that those individuals in clusters randomly allocated to control have always control treatment, but those in clusters randomised to the active intervention can switch to the control treatment, depending on their adherence class. We finally generate continuous outcome Y_{ij}, under the exclusion restriction assumption

$$Y_{ij} = \beta_0 + \beta_C C_{ij} + \beta_{CZ} C_{ij} Z_j + \beta_W W_j + \beta_X X_{ij} + \upsilon_j + \epsilon_{ij} \tag{10}$$

with $\upsilon_j \sim N(0, \sigma_\upsilon^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where the values for σ_υ^2 and σ_ϵ^2 are chosen, such that the marginal ICC for Y has the corresponding value according to the simulated scenario, given that $\text{Var}(Y_{ij}) = \sigma^2 = 1$.

For simplicity, but without loss of generality, we assume that there is no direct effect of complying on the outcome, and thus $\beta_C = 0$, so that the mean potential outcome in the control never-takers is equal to the mean potential outcome of the control compliers and thus complying with the control treatment has no effect on the outcome. Since $\beta_{CZ} \neq 0$, there is a non-zero effect of complying with the active arm (i.e. receiving active treatment). The choice of β_C does not affect our estimation, as the effect of β_C (and the intercept β_0) cancels out for the average treatment effect in the compliers and non-compliers, respectively.

The choice of the parameters' values is reported in Table 1.

Table 1. Factorial design of the data generating processes and values taken by the parameters in the simulations.

Parameter	Label	Level	Value
CRT size			
N	Total number of individuals	Moderate	≈ 1000
J	Number of clusters and	Moderate clusters	$J = 50, n_j \sim \text{Poi}(20)$
n_j	individuals per cluster	Few large clusters	$J = 10, n_j \sim \text{Poi}(100)$
Baseline variables			
W_j	CL variable	–	$W_j \sim N(0, 0.08)$
ρ_X	ICC for X_{ij}	Moderate	0.05
X_{ij}	Individual-level variable	–	$X_{ij} = X_j + e_{ij}, X_j \sim N(0, 0.004)$, $e_{ij} \sim N(0, 0.076)$
Adherence to treatment			
	Expected probability of adherence	Moderate	0.60 (CL adherence) 0.85 (individual-level adherence)
λ_W, λ_X	W_j and X_{ij} effects on log odds of adherence	Small	$\lambda_W = 0.05, \lambda_X = 0.05$
		Large	$\lambda_W = 0.70, \lambda_X = 0.70$
C_j	CL adherence class	–	Bern [expit$(\lambda_0 + \lambda_W W_j)$]
C_{ij}	Individual-level adherence class	–	Bern [expit$(\lambda_0 + \lambda_W W_j + \lambda_X X_{ij} + \zeta_j)$]
ζ_j	CL random effects	–	$\zeta_j \sim N(0, \pi^2/3)$
ρ_C	ICC for C_{ij}	Moderate	0.50
Outcome			
β_0	Intercept		$\beta_0 = 0$
β_C	Effect of complying amongst controls		$\beta_C = 0$
β_W, β_X	W_j and X_{ij} effects on outcome Y_{ij}	Small	$\beta_W = 0.1\ \sigma$, $\beta_X = 0.1\ \sigma$
		Large	$\beta_W = 0.4\ \sigma$, $\beta_X = 0.4\ \sigma$
β_{CZ}	True LATE	Small, Large	$0.1\ \sigma$, $0.4\ \sigma$
ρ_Y	ICC for Y_{ij}	Small, Large	0.05, 0.20

Table 2. Overview of TSLS estimation and inference strategies used in the simulation study.

Analyses strategy	Levels	
CL outcome	Unadjusted	Adjusted for X_{ij}
TSLS adjusted for W_j	No	Yes
Least square method	Ordinary	Weighted
Weights (if using)	CS	MV
SE estimation	Normal theory	HW SE
SSDF correction	No	Yes

CL: cluster level; HW: Huber–White; CS weights: cluster-size weights; MV: minimum variance; SE: standard error; SSDF: small sample degrees of freedom correction.

We need the data generating process to result in randomised treatment Z being a valid IV, but with this choice of parameters, some combinations may result in weak instruments, for example, CL non-adherence settings, with only five clusters per arm, and the proportion of non-adherent clusters set at 40% (the median proportion of non-adherent clusters reported in Agbla and DiazOrdaz[4] being 44.8%). Thus, after creating each dataset, we perform an unadjusted first stage regression of D_j on Z_j and reject simulated datasets where the resulting $F-$statistic is <10 (Staiger and Stock's rule of thumb for weak instruments[36]). We continue this process until we have 2500 datasets per scenario.

Estimation in each scenario involves using unCL summary of treatment received in the first stage, and either unadjusted or individual-level variable adCL summary outcomes, for the second stage. Each regression in the TSLS was fitted via ordinary least squares or generalised least squares, the latter with either cluster size or minimum-variance weights. We also consider TSLS where each stage model is either unadjusted or adjusted for a CL variable. Finally, we obtain SEs assuming homoscedasticity or Huber–White SEs and SSDF-based or normal approximation CIs. A summary is given in Table 2. Details of the Stata code used for analysis are found in the online Appendix.

The performance criteria used are empirical bias and coverage rates of the 95% CIs over the 2500 replicate datasets per scenario. For the bias, we construct a 95% CI using the Monte Carlo errors. The coverage rate sampling error, given the size of the simulation, results in a valid range between 94.1 and 95.9%. See Appendix 2 for the formal definitions.

Results

We present the results by plotting the empirical bias with the Monte Carlo error-based CIs. The coverage rate valid range is represented by horizontal-dashed lines.

Figures 1 and 2 report the empirical bias and 95% CI coverage resulting from each of the different CL–TSLS estimators, when adherence is at CL or individual level, respectively, and for scenarios where the true LATE is large. The corresponding figures for small true LATE are shown in Appendix 4 (Figures 7 and 8).

Each figure reports results where $J = 10$ (Panel A, top) or $J = 50$ (Panel B), and with the ICC for Y, ρ_Y is either small (first three columns) or large (last three columns). In each cell, the results for alternative combinations of TSLS (unadjusted/adjusted for W_j) applied to unCL or adCL outcomes are plotted along the horizontal axis. The different data generation scenarios are identified by $*$, $+$, $\times$ and $^\circ$, corresponding to varying strengths of the effects of X and W on Y.

We see that all CL–TSLS estimators show some finite sample bias in settings where the number of clusters is small ($J = 10$, Panel A), regardless of whether the non-adherence was at the cluster or individual level and whether the CL summary for Y was adjusted or unadjusted or W_j was included or not in the TSLS regressions. However, the Monte-Carlo error CIs includes 0 in many settings. The bias is more severe when the ICC for Y is larger (right hand side of each figure). The bias is somewhat attenuated when we adjust for W_j in the TSLS, and the non-adherence is at the CL (Figures 1 and 7). In contrast, for settings with individual-level non-adherence, this adjustment instead increases the bias, especially if W has only a small confounding effect. In these scenarios, the estimates exhibit a small but statistically significant bias, which disappears when the number of clusters is larger (Figures 2 and 8). In general, the bias is not affected by the choice of weighting strategy nor by whether ρ_Y is small or large. The bias is negligible for settings where the number of clusters is moderate or large ($J = 50$).

Comparing the results of the second, third and fourth rows in each panel (Figures 1 and 2), we see that the coverage rate is affected by the choice of SE estimation and also by whether SSDF correction is used. When the

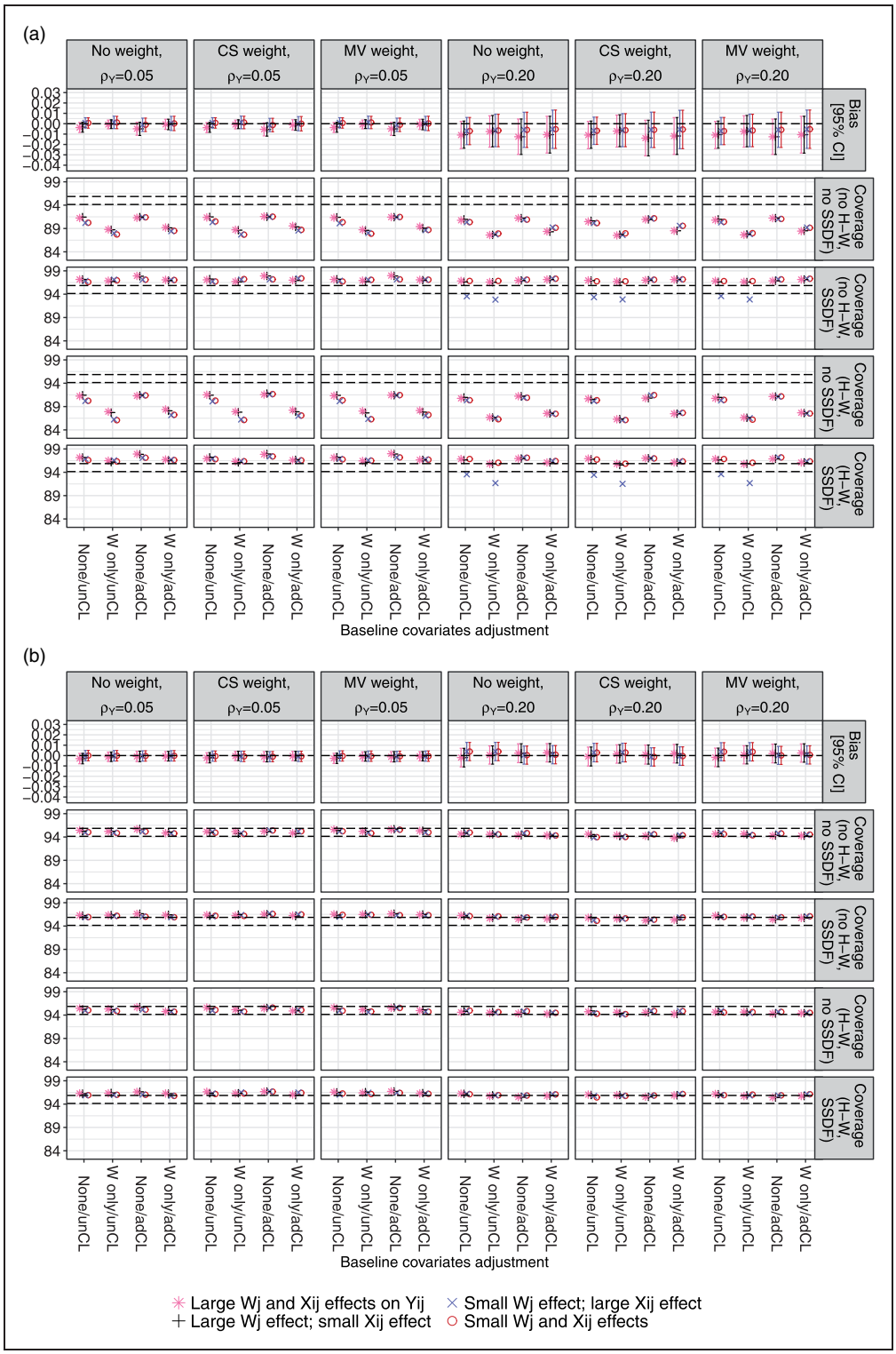


Figure 1. Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with CL non-adherence and large true LATE. Data generation scenarios represented by $*$, $+$, $\times$, and $\circ$. Estimates are obtained via unadjusted or *W*-adjusted TSLS with different weights (none, cluster size (CS), and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for *X* outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panels A and B.

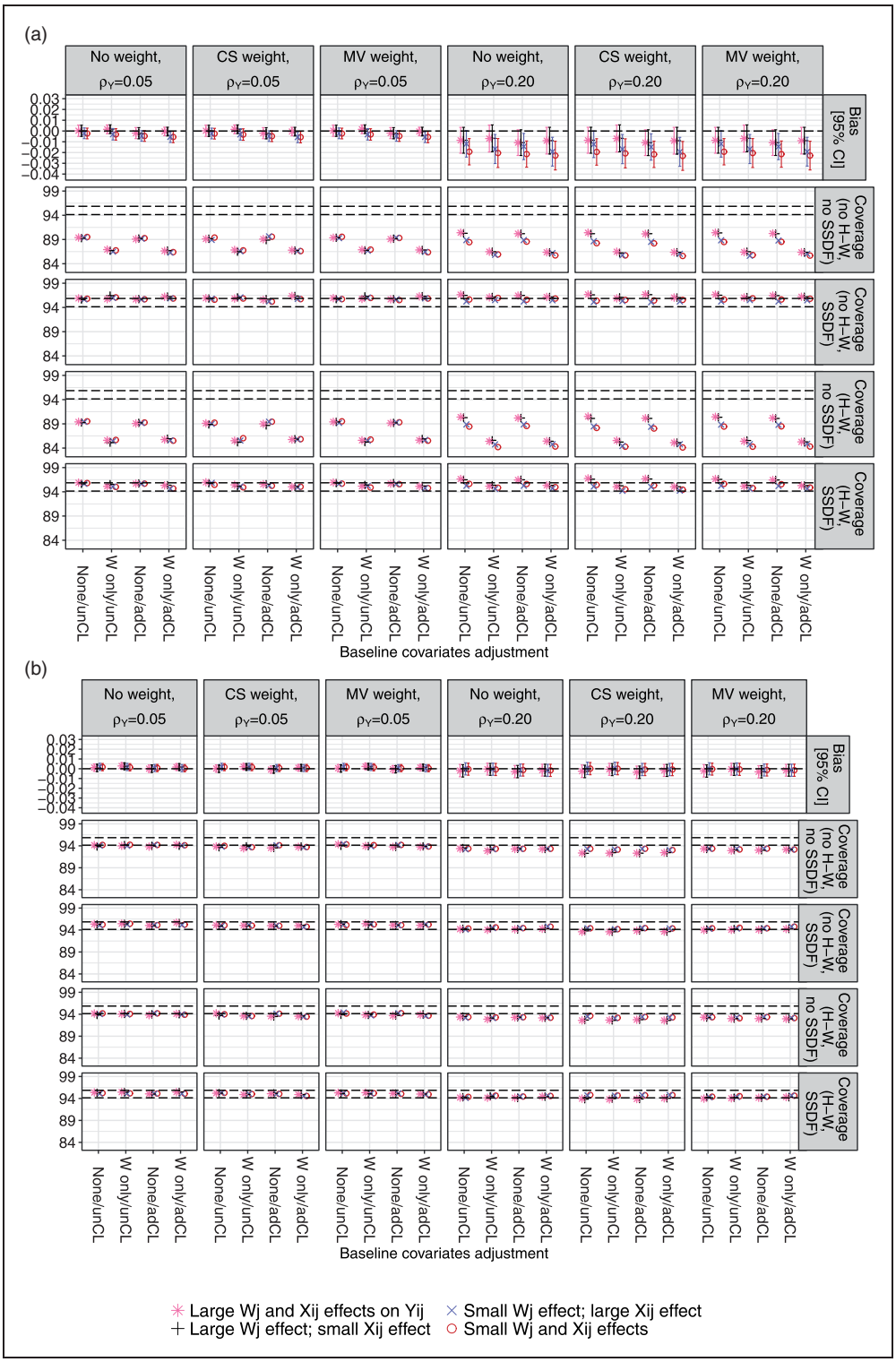


Figure 2. Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and large true LATE. Data generation scenarios represented by $*$, $+$, $\times$, and $^\circ$. Estimates are obtained via unadjusted or *W*-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for *X* outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panels A and B.

number of clusters is small, an SSDF correction must be used as failing to do so results in under-coverage (Panel A). The low coverage is more serious when TSLS adjusts for W (second and fourth set of results in each panel).

Overall, the results in Panel A of each figure show that coverage is closer to the nominal levels when using SSDF correction when constructing CIs (third and fifth rows). Using Huber–White SE or not has little to no impact if there is no SSDF correction. However, the SSDF correction for the CIs resulting from a TSLS using unCL outcomes can lead to under-coverage, as shown in the specific case with CL non-adherence, large ρ_Y and large true LATE, but where only X is strongly associated with Y (Figures 1 and 7, third and fifth rows of Panel A, right-hand side columns, scenario represented by $\times$ in the plots). The use of adCL outcomes (i.e. where the CL outcome is the residual after adjusting for individual-level variable X) recovers coverage to values close to nominal. This is not the case when the non-adherence is at the individual level, and both W and X are confounders in the data generating process.

In both cluster and individual-level non-adherence settings, it can be seen that using minimum-variance weights increases the coverage by a small fraction, when compared with cluster size weights, especially for scenarios with $J = 50$ and large ρ_Y. However, since minimum-variance weights require an estimate of the CL variance, and this is badly estimated when the number of clusters is small ($J = 10$), we can see that the minimum-variance weights are less efficient than using either no weights or cluster-size weights. This is most clearly seen when no Huber–White SE correction has been used.

We can also see that when SSDF correction is used, then not using Huber–White SE can result in small over-coverage especially for CL non-adherence settings, which is improved when Huber–White SE are used (Figures 1 and 7, third and fifth rows of Panel A). When $J = 50$ (Panel B), the use of SSDF-based distributions is not expected to make any material difference, and this is indeed the case. The impact of using Huber–White SE or the different weighting strategies is also minimal.

Additional simulations

Two extra additional scenarios are now considered to investigate the sensitivity of the CL–TSLS performance to number of clusters and cluster-size imbalances, at both CL and individual-level adherence, but focusing on settings where confounding is strong with a large true LATE.

In the first additional simulation, we explore the impact that the outcome ICC and the number of clusters have on bias, while leaving the expected total sample size fixed (= 1000).

We consider two marginal ICC for Y_{ij} ($\rho_Y = 0.05$ and $\rho_Y = 0.80$) and three average cluster sizes ($n_j = 20$, 10 and 2.5, corresponding to whether the number of clusters varied from $J = 50$, 100 or 400), which includes one of the scenarios previously considered in the main simulations for comparison. Although CRTs rarely have ICCs above 0.10,[37] the value of $\rho_Y = 0.80$ is included to evaluate the performance of the methods in extreme settings.

In the second additional set of simulations, we explore the effect of high cluster-size imbalances. While keeping the average sample size equal to 1000, and $J = 10$ or 50, we create high cluster size imbalance using a Pareto distribution to generate the cluster sizes.[38] The Pareto distribution parameters are chosen so that approximately 40% of the clusters have a size below 15, and 60% a size above 15, while the average cluster size is 20 and the minimum cluster size is 10, resulting in approximately 1.8 for the shape and 9.1 for the scale.

Results

Figures 3 and 4, corresponding to CL and individual-level non-adherence settings, show that for a fixed number of clusters (cells in the same row), the bias increases with increasing ICC for Y, but that as the number of clusters increases (moving down the column in the figure), CL–TSLS results in negligible mean bias, even a very large ρ_Y. It is well known that TSLS is only asymptotically unbiased, and with CL analyses, we expect the asymptotics to depend on the number of clusters and not the number of individuals. Nevertheless, the CL-summaries treated as outcomes for the two models involved in TSLS contain less 'information' when the ICC is higher, which translates into a larger number of clusters being necessary for the bias to be negligible.

The impact of high cluster size imbalance is reported in Figures 5 and 6, where non-adherence is at the CL and individual level, respectively. We see that even with the use of Huber–White SE, failure to do SSDF correction results in under-coverage, especially when ρ_Y is large. Looking at Panel B in Figure 6, we can see that using cluster-size weights results in even lower coverage. This is because cluster-size weights are known to perform well when the CL residuals are homoscedastic, which is unlikely when cluster sizes are very imbalanced.[15] The use of SSDF correction brings the coverage close to nominal levels.

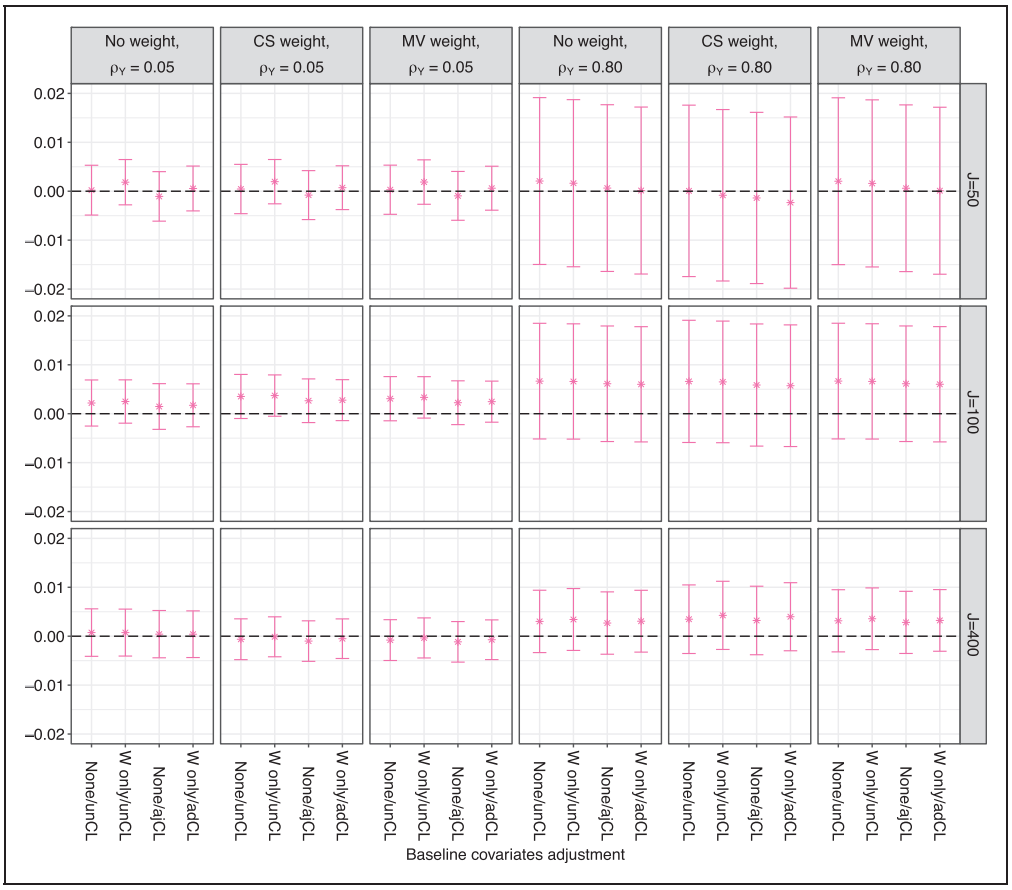


Figure 3. Bias of the CL-LATE for the extra simulation where non-adherence is at the CL and a large true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

Illustrative example

We now illustrate the methods in practice by applying each in turn to the analysis of the TXT4FLUJAB trial. This was a CRT of general practices (GP) in England aiming at estimating the effect of text messaging influenza vaccine reminders on increasing vaccine uptake in patients with chronic conditions, carried during the 2013 influenza season.[20] GPs were stratified by the type of software used for text messaging and randomised to either standard care (control group, 79 GPs and 51,136 patients) or a text messaging campaign (active group, 77 GPs and 51,121 patients). Practices were not blinded to their allocation. GPs were the unit of analysis, and the outcome of interest was the proportion of influenza vaccine uptake at the GP level.

Influenza vaccination within the GPs was automatically recorded in the clinical system from which the data were extracted, so there are no missing data.

Since non-adherence was anticipated, the original statistical analysis plan specified obtaining by IV regression an efficacy estimate at the GP level.[20] The original publication reported an estimated increase in vaccine uptake from texting reminders of 14.3% (95% CI: –0.59%–29.2%),[20] after dichotomising adherence at the CL as either 100% of eligible patients, compared with texting $<100\%$.

Adherence to the intervention at the individual level could not be measured for all practices because it was recorded in a usable form only for GPs using a specific software. Therefore, for these re-analyses, we restrict the dataset to 116 GPs (58 in the intervention and 58 in the standard care arm) for which individual-level adherence data are available. Six of the 58 practices (10%) in the intervention arm did not send any reminders. Conversely, 21 of the 58 practices (36%) in the standard care arm actually sent a reminder to at least one patient. Hence non-adherence is two-sided. It also varies at the individual level. The median (range) of percentage of non-adherence at the GP level was 0 (0–78.4%) and 21.0% (0–83.5%) in the control and active group, respectively (Table 3).

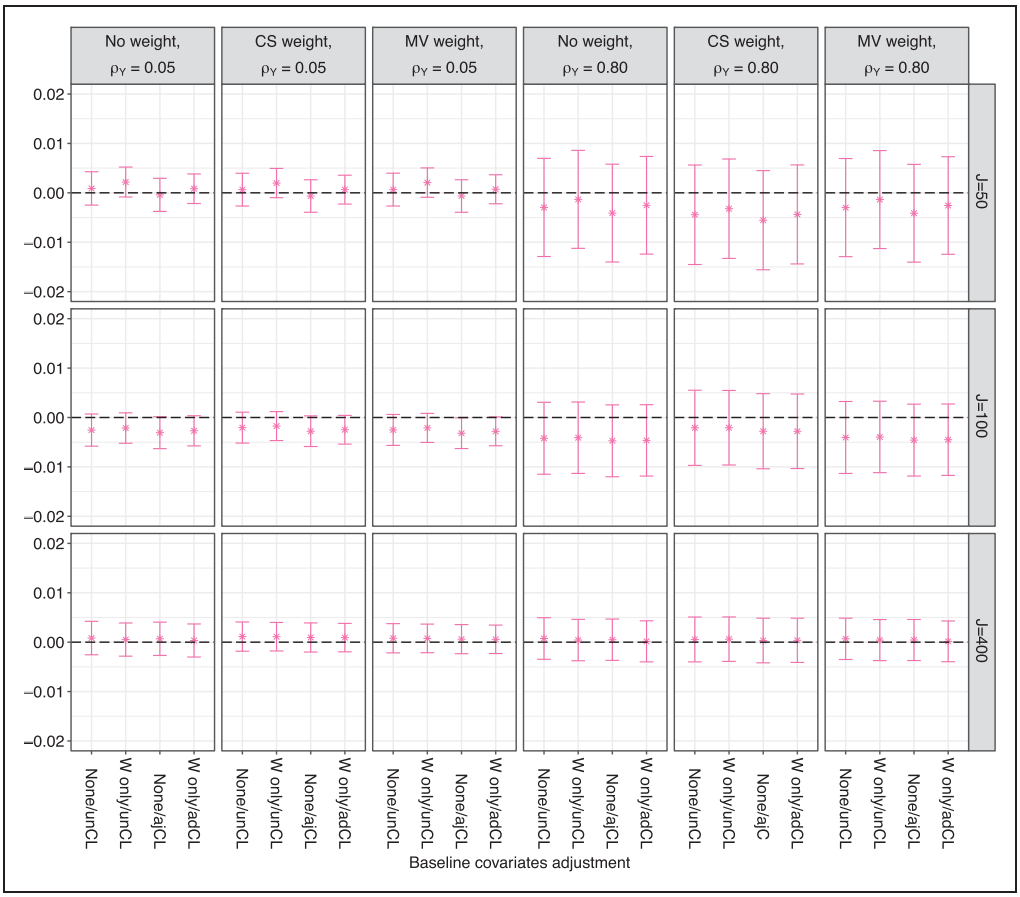


Figure 4. Bias of the CL-LATE for the extra simulation where non-adherence is at the individual-level and a large true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

The characteristics of the GPs and of the patients included in these analyses are comparable across trial groups (Table 3); further, the marginal ICC for individual-level outcome (vaccination) and treatment received (text message reminder) was 0.03 and 0.84 on the log-odds scale, respectively.

For our re-analysis, we begin by discussing the plausibility of the necessary assumptions.

First, the consistency assumption in this setting implies that the means of sending and receiving the text message, as well as the timings are irrelevant, in the sense that all of these would lead to the same observed outcome. So whether the text was pre-programmed, or sent by a doctor or a nurse, or received in the morning or at night or weekend, it would have the same effect of either getting the patient to be vaccinated, or not, irrespective of any of these factors. This seems a reasonable assumption for this intervention.

In contrast, there is a small risk of interference. The cluster defined by GP practice should minimise this, as we only need to assume no interference at the CL, but it could be plausible that patients interact with those outside their GP, so that the exposure to a text message reminder of one patient may indeed affect the potential outcome, in this case, influenza vaccination of another patient from a different GP. The risk is small as usually close family members belong to the same GP.

Now, regarding the identification assumptions, we note that the unconfoundedness of the CL randomised treatment assumption is satisfied by design. To check whether cluster randomisation is a relevant instrument, we perform a test on the first stage of the CL–TSLS. The corresponding F-statistic is $F(1, 114) = 28.7 > 10$, thus passing Staiger and Stock's rule of no null first-stage.[36]

The exclusion restriction at the individual level implies that there is no other mechanism by which the GP being randomised to sending text vaccination reminders can affect a patient's actual vaccination uptake beside via the sending of the message. This assumption needs further justification, as in principle, a GP randomised to send reminders can be more conscious of the risks the patients face during the influenza season and use other means to

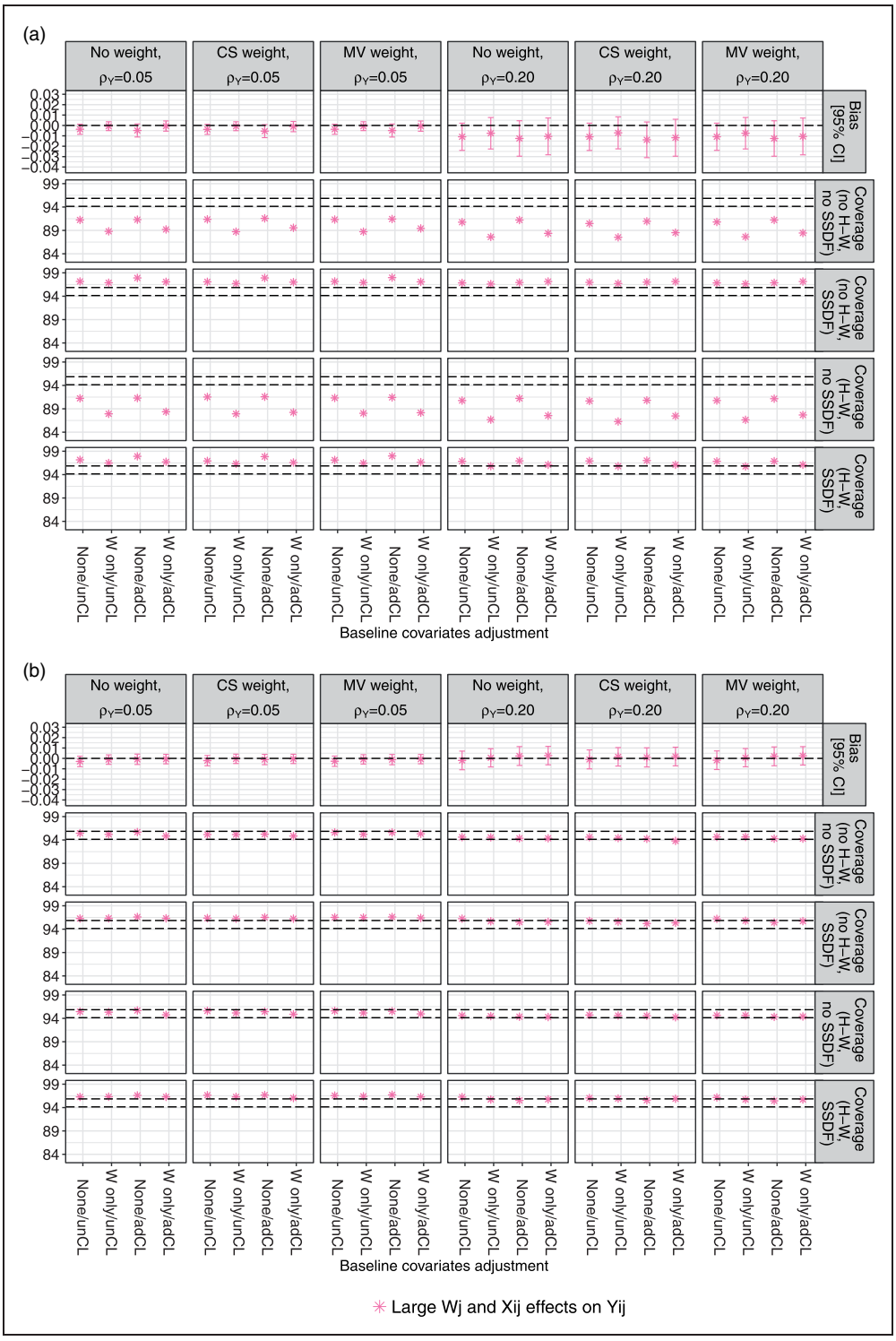


Figure 5. Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber–White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the CL, and a large true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS), and minimum-variance (MV)). Small and large number of clusters results appears in Panels A and B, respectively.

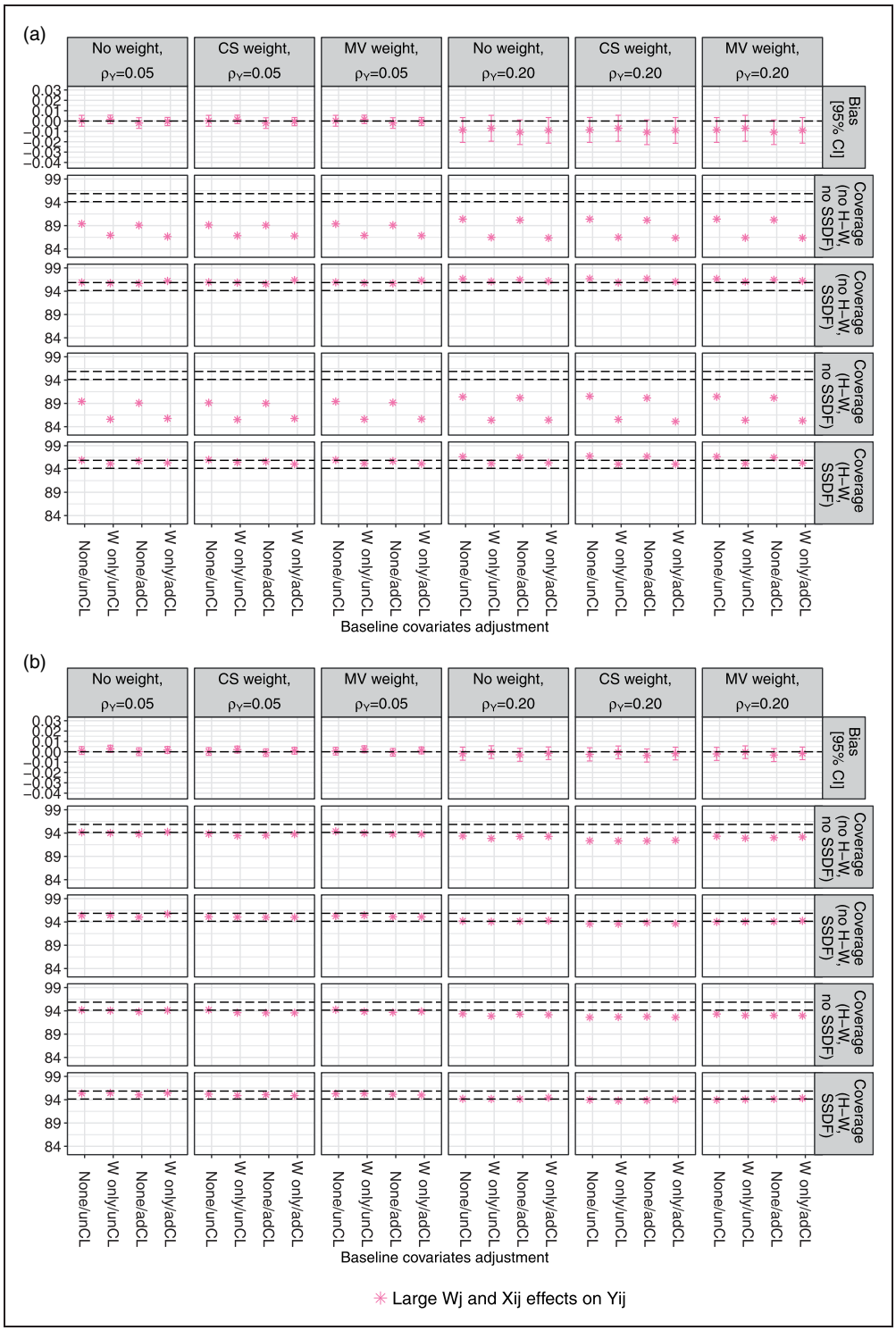


Figure 6. Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber–White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the individual-level, and a large true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS), and minimum-variance (MV)). Small and large number of clusters results appears in Panels A and B, respectively.

remind at-risk patients, either in person, by post or by putting out flyers and posters in the clinic. So, it is possible that there are patients who do not receive text reminders and yet are prompted to get vaccinated by other means, by virtue of their practice being in the active group. However, flyers, posters and postal letters already form part of regular care, so we believe they do not really vary by whether the GP is randomised to the active group.

Finally, the monotonicity assumption (that there are no defiers) also seems plausible as GPs randomised to the active group were more likely to send a text message reminder than those in the control group (see Table 3).

CL–TSLS on unCL outcomes was implemented adjusting and not adjusting for a baseline CL covariate, namely whether the clinic was open on the weekends (yes/no). Table 4 shows the CL–LATE estimates (expressed as mean risk differences), with 95% CIs and p-values obtained via different weighting strategies and corrections.

Using cluster-size weights results in different point estimates from the rest. This was expected as there is substantial cluster-size imbalance (cluster size range: 148–1678 in the control group and 79–3022 in the active group (Table 3)). The results obtained using no weights or minimum-variance weights lead to point estimates that are very close to those found in the original publication.[20]

Table 3. Baseline characteristics and percentages of non-adherence for the TXT4FLUJAB trial.

Characteristics	Control	Active
Practice-level characteristics		
Number of practices, *n* (%)	58 (100.0)	58 (100.0)
Open on weekends, *n* (%)	39 (67.2)	37 (63.8)
Patients per practice, median (range)	660 (148–1678)	684 (79–3022)
Patient-level characteristics		
Number of patients, n (%)	40633 (100)	41073 (100)
Male, n (%)	20 752 (51.1)	21 012 (51.2)
Has any disease, n (%)	39244 (96.6)	39672 (96.6)
Age, median (range)	50 (18–64)	50 (18–64)
Active treatment received		
Patients receiving text message reminders, *n* (%)	2628 (6.5)	11113 (27.1)
Practices sending text message reminders, *n* (%)	21 (36.2)	52 (80.7)
% of patients in each GP receiving reminders, median (range)	0 (0–78.4)	21.0 (0–83.5)

Table 4. LATE of text message reminders to receive flu vaccination on the uptake of flu vaccine in the TXT4FLUJAB trial, using unadjusted CL-summaries.

		Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p
No weighting	None	0.149 (−0.006, 0.305)	0.060	0.148 (−0.078, 0.303)	0.063
	HW	0.149 (−0.006, 0.305)	0.060	0.148 (−0.005, 0.301)	0.058
	SSDF	0.149 (−0.009, 0.308)	0.065	0.148 (−0.012, 0.308)	0.069
	SSDF + HW	0.149 (−0.009, 0.308)	0.065	0.148 (−0.009, 0.305)	0.064
Cluster size weights	None	0.071 (−0.065, 0.207)	0.307	0.074 (−0.061, 0.209)	0.284
	HW	0.071 (−0.088, 0.230)	0.382	0.074 (−0.077, 0.225)	0.338
	SSDF	0.071 (−0.068, 0.209)	0.313	0.074 (−0.064, 0.212)	0.292
	SSDF + HW	0.071 (−0.091, 0.233)	0.388	0.074 (−0.081, 0.228)	0.346
Minimum-variance weights	None	0.143 (−0.008, 0.293)	0.064	0.142 (−0.009, 0.293)	0.065
	HW	0.143 (−0.006, 0.291)	0.060	0.142 (−0.005, 0.289)	0.058
	SSDF	0.143 (−0.011, 0.296)	0.069	0.142 (−0.012, 0.297)	0.071
	SSDF + HW	0.143 (−0.009, 0.294)	0.065	0.142 (−0.008, 0.293)	0.064
Weighting strategy	SE and correction	Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p

[a]Adjusted for whether clinic is opened during weekends.
HW: Huber-White; SSDF: small sample degrees of freedom.

In terms of inference, the use of SSDF correction in calculating CIs is not important, as the number of clusters is large, but the Huber–White SEs paired with minimum-variance weighting provides efficiency gains, especially for the adCL–TSLS analyses. Overall, however, the CIs are still very wide.

These results suggest that there is some evidence that receiving a text reminder increases the expected proportion of patients within a compliant practice that get vaccinated against influenza by 14% (95% CI: –0.5 to 29.3%, $p = 0.058$, based on the adCL–TSLS using minimum-variance weights and normal-based CI with Huber–White SEs estimate).

Contrast this with the unCL-summaries mean risk difference ITT estimate, which indicates a 2.89% increase (95% CI: –0.17–5.95, $p = 0.064$), highlighting the dilution effects deriving from the non-adherence.

One of the disadvantages of TSLS is lack of efficiency. Adjusting for individual-level baseline covariates may help obtaining narrower CIs. Since CL–TSLS cannot adjust for individual-level covariates, we now perform the analyses using adCL summary outcomes, generated by adjusting for gender, age and the presence of disease. Results are reported in Table 6 in Appendix 3. The results do not materially change (weak evidence of a 13% increase vaccination uptake), possibly because these individual-level covariates are not strongly associated with the outcome.

The illustrative example shows the importance of choosing and pre-specifying the TSLS analysis according to the trial characteristics. In the example, the choice of weights changed the point estimate to such an extent that the small evidence in support of treatment benefit disappeared completely. So, if the trial has very imbalanced cluster sizes, Huber–White corrections can help for the SEs, but the point estimates may be biased, if large clusters are somewhat atypical.

Our application is limited by the availability of baseline CL variables. Since there was only one CL variable recorded, the impact of covariate adjustment on the CL–TSLS is negligible. Other limitations of these results include the possibility of measurement error, for if patients received their influenza vaccine outside the practice, this would not have been recorded in the system, unless the patient informed their GP.

Discussion

This paper demonstrates the use of TSLS regression applied to CL summaries (CL–TSLS) as a simple and valid method for obtaining estimates of the LATE in CRTs where non-adherence occurs at either the CL or the individual level. To improve the efficiency of CL–TSLS estimates, we proposed adjusting for baseline variables; if these are CL, in the TSLS regression, while if these are individual level, by adjusting the CL-summary outcomes before performing TSLS. The performance of CL–TSLS regression of either adjusted or unCL-outcomes and adjusting or not for CL-baseline variables was evaluated with different weighting strategies (none, cluster size, minimum variance) as well as the use of different methods for constructing CIs (alternatively using or not Huber–White SEs and/or SSDF correction) in a factorial simulation study.

We have demonstrated empirically through simulations that under the stated sufficient assumptions for identification, TSLS regression of CL summaries provides consistent estimates of the causal treatment effect in the sub-population of compliers, where non-adherence is at the CL. With individual-level non-adherence, the additional assumption that the cluster-specific LATE is the same across clusters is required for CL–TSLS to identify the population LATE.[31] Moreover, provided that an appropriate distribution with SSDF adjustment is used when the number of clusters is small and Huber–White SEs are used if there is high cluster size imbalance, valid 95% CIs can be constructed.

Our simulation study suggests that all weighting strategies perform similarly when the number of clusters is not small. When the number of clusters is small, minimum-variance weights tend to be badly estimated and are not recommended; furthermore when the cluster sizes are very variable, cluster-size weights should not be used. Although in the simulations, the choice of weights did not affect the point estimates, and these were affected in the illustrative example. Overall, our results show that, unless there are very few clusters, or the outcome ICC is large, minimum-variance weighting performs well.[23] An overall summary of these findings in the format of recommendations is given in Table 5.

Although CL–TSLS is easy to implement, it suffers from being very inefficient. We can see this in the illustrative example where all CIs for the CL–TSLS LATE estimates are much wider than those for the estimated ITT. There are two reasons for this: CL analyses are inefficient, unless the cluster sizes are (almost) equal,[39] and TSLS is known to be inefficient, although adjusting for baseline covariates can ameliorate this.[14] In the context of CL analysis, it is only possible to include CL baseline covariates in the regressions.[15] However, we tested the performance of CL outcomes which are adjusted for individual-level covariates[26] and showed that this indeed has the potential to improve efficiency in certain settings.

Table 5. Recommendations.

Adherence	Comments
At CL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	Adjust for CL variables in TSLS to reduce bias and improve efficiency
If an IL variable is a strong confounder	Use adjusted CL-outcomes in the TSLS to improve efficiency
If CS is imbalanced	Use small sample DF correction to improve inference and avoid using CS weights
At IL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	avoid adjusting for CL variables
If CS are imbalanced	Use small sample DF correction to improve inference and avoid using CS weights

CS: cluster sizes; CL: cluster level; DF: degrees of freedom; IL: individual level.

For CL–TSLS analyses, inference should be based on the number of clusters, with CIs constructed by using $t-$distributions with degrees of freedom equal to $J - p$.[40] The outcome ICC value is important too, with higher ICCs requiring a larger number of clusters for the asymptotical arguments to work as well as whether the CL variances are homoscedastic.[15]

Other methods for estimating causal treatment effects in CRTs with non-adherence at the individual level exist, in particular, Kang and Keele[31] have recently proposed a finite-sample estimator that identifies the population LATE and obtains valid inferences even when compliance is low.

We do not consider here situations where the identification assumptions are violated. There are several options to study the sensitivity to departures from these assumptions. For example, if the exclusion restriction does not hold, a Bayesian parametric model can use priors on the non-zero direct effect of randomisation on the outcome for identification.[41] Since the models are only weakly identified, the results depend strongly on prior distributions. Alternatively, violations of the exclusion restriction can also be handled by using baseline covariates to model the probability of compliance directly, within structural equation modelling via expectation-maximisation.[42,43]

We have only focused on LATE estimands. These are often criticised because the estimates obtained apply to the 'compliers' in the population, and these cannot be observed in practice, thus limiting applicability. However, LATE estimates may be used to provide information about the average causal effect in the entire population.[44] Moreover, the average treatment effect on the compliers is often of interest to patients and medical decision makers, especially when they expect patients to comply with the treatment.[45]

Acknowledgement

We thank the TXT4FLUJAB study team for access to the data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The authors received financial support from a UK Economic and Social Research Council PhD scholarship ES/J500021/1 and UK Medical Research Council Career development award in Biostatistics MR/L011964/1 for the research, authorship, and/or publication of this article.

ORCID iD

Karla DiazOrdaz https://orcid.org/0000-0003-3155-1561

References

1. Donner A and Klar N. Pitfalls of and controversies in cluster randomization trials. *Am J Publ Health* 2004; **94**(3): 416–422.
2. Glynn RJ, Brookhart MA, Stedman M, et al. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care* 2007; **45**(10): S38–S43.
3. Wright N, Ivers N, Eldridge S, et al. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015; **68**(6): 603–609.
4. Agbla SC and DiazOrdaz K. Reporting non-adherence in cluster randomised trials: a systematic review. *Clin Trial* 2018; **15**(3): 294–304.
5. Schochet PZ and Chiang HS. Estimation and identification of the complier average causal effect parameter in education RCTs. *J Educ Behav Stat* 2011; **36**(3): 307–345.
6. Hernán MA and Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med* 2017; **377**(14): 1391–1398.
7. Akacha M, Bretz F and Ruberg S. Estimands in clinical trials: broadening the perspective. *Stat Med* 2017; **36**(1): 5–19.
8. Imbens GW and Angrist JD. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62**(2): 467–475.
9. Angrist JD, Imbens GW and Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; **91**(434): 444–455.
10. White IR. Uses and limitations of randomization-based efficacy estimators. *Stat Methods Med Res* 2005; **14**(4): 327–347.
11. Bellamy SL, Lin JY and Ten Have TR. An introduction to causal modeling in clinical trials. *Clin Trial* 2007; **4**(1): 58–73.
12. Angrist JD and Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J Am Stat Assoc* 1995; **90**(430): 431–442.
13. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: *Proceedings of the American statistical association, section on Bayesian statistical science*, Amer. Statist. Soc., Alexandria, VA, 2000, pp. 6–10.
14. Wooldridge JM. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press, 2010.
15. Angrist JD and Pischke JS. *Mostly harmless econometrics: an empiricist's companion*. Princeton: Princeton University Press, 2008.
16. Frangakis CE, Rubin DB and Zhou XH. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* 2002; **3**(2): 147–164.
17. Jo B, Asparouhov T, Muthén BO, et al. Cluster randomized trials with treatment noncompliance. *Psychol Meth* 2008; **13**(1): 1.
18. Schochet PZ. Estimators for clustered education RCTs using the Neyman model for causal inference. *J Educ Behav Stat* 2013; **38**(3): 219–238.
19. Campbell MK, Mollison J, Steen N, et al. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract* 2000; **17**(2): 192–196.
20. Herrett E, Williamson E, van Staa T, et al. Text messaging reminders for influenza vaccine in primary care: a cluster randomised controlled trial (TXT4FLUJAB). *BMJ* 2016; **6**(2): e010069.
21. Campbell MJ and Walters SJ. *How to design, analyse and report cluster randomised trials in medicine and health related research? Statistics in practice*. New York: Wiley, 2014.
22. Prais SJ and Aitchison J. The grouping of observations in regression analysis. *Rev Inst Int Stat* 1954; **22**: 1–22.
23. Kerry SM and Bland MJ. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Stat Med* 2001; **20**(3): 377–390.
24. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**(4): 817–838.
25. Cameron AC and Trivedi PK. *Microeconometrics: methods and applications*. Cambridge: Cambridge University Press, 2005.
26. Hayes R and Moulton L. *Cluster randomised trials*. London: Taylor & Francis, 2009.
27. Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J Am Stat Assoc* 2006; **101**(476): 1398–1407.
28. Vander Weele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; **20**(6): 880–883.
29. Cole SR and Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 2009; **20**(1): 3–5.
30. Imbens GW and Rubin DB. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge: Cambridge University Press, 2015.
31. Kang H and Keele L. Estimation methods for cluster randomized trials with noncompliance: a study of a biometric smartcard payment system in India. *arXiv Prepr* arXiv:1805.03744v2.
32. Hernán MA and Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**: 360–372.
33. Swanson SA and Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity. *Int J Epidemiol* 2017; **47**(4): 1289–1297.
34. Vansteelandt S and Didelez V. Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. *Scand J Stat*, Epub ahead of print 24 May 2018. DOI: 10.1111/sjos.12329).

35. White H. Instrumental variables regression with independent observations. *Econometrica* 1982; **50**: 483–499.
36. Staiger DO and Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997; **65**(3): 557–586.
37. Murray DM and Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev* 2003; **27**(1): 79–103.
38. Guittet L, Ravaud P and Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. *BMC Med Res Methodol* 2006; **6**(1): 17.
39. Donner A. Some aspects of the design and analysis of cluster randomization trials. *J R Stat Soc Ser C (Appl Stat)* 1998; **47**(1): 95–113.
40. Donald SG and Lang K. Inference with difference-in-differences and other panel data. *Rev Econom Stat* 2007; **89**(2): 221–233.
41. Conley TG, Hansen CB and Rossi PE. Plausibly exogenous. *Rev Econom Stat* 2012; **94**(1): 260–272.
42. Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Stat Med* 2002; **21**(21): 3161–3181.
43. Jo B. Estimation of intervention effects with noncompliance: alternative model specifications. *J Educ Behav Stat* 2002; **27**(4): 385–409.
44. Baiocchi M, Cheng J and Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014; **33**(13): 2297–2340.
45. Murray EJ, Caniglia EC, Swanson SA, et al. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. *J Clin Epidemiol* 2018; **103**: 10–21.

Appendix 1. Adjusted CL summaries for binary data

For binary, a standard logistic regression model is usually fitted for binary outcomes, which assumes that

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \lambda_1 + \lambda_2 X_{ij} \tag{11}$$

Let M_j and $\hat{M}_j$ be the observed and predicted number of successes in the jth cluster, respectively. After fitting model (11), $\hat{M}_j$ is calculated as

$$\hat{M}_j = \sum_{l=1}^{m} \hat{\pi}_{ij} = \sum_{i=1}^{n_j} \text{expit}\left(\hat{\lambda}_1 + \hat{\lambda}_2 X_{ij}\right)$$

Then the observed and predicted numbers of success are compared by computing a residual for each cluster. If we want to estimate the adjusted RD, the residual, known as difference-residual, for each cluster is calculated as

$$e_j = (M_j - \hat{M}_j)/n_j$$

and treated as a continuous outcome in any subsequent analyses.

Appendix 2. Performance criteria

Let the mean of the estimated LATE across the replicate datasets in each scenario, indexed by $l = 1, \ldots, L$, with $L = 2500$ be $\bar{\hat{\beta}}_{IV} = \frac{1}{L}\sum_{l=1}^{L} \hat{\beta}_{IV_l}$. The following criteria were used to assess the performance of the methods investigated.

(a) Empirical bias: estimated by $\bar{\hat{\beta}}_{IV} - \beta_{CZ}$.

(b) Monte Carlo error of empirical bias $= \sqrt{\sum_{l=1}^{L}\left(\hat{\beta}_{IV_l} - \bar{\hat{\beta}}_{IV}\right)^2/[L(L-1)]}$.

(c) Coverage rate of the nominal of 95% CIs $\frac{1}{L}\sum_{l=1}^{L} I(|\hat{\beta}_{IV_l} - \beta_{CZ}| < 1.96 s_i)$, where s_i denotes the model-based SE for $\hat{\beta}_{IV_l}$. The Monte Carlo Error of coverage is $\sqrt{\sum_{l=1}^{L}(0.95)(0.05)/L}$.

Appendix 3. Results for adjusted CL summaries CL–TSLS for TEXT4FLUJAB

Table 6. TSLS estimation of practice-level LATE of reminder text messaging to receive flu vaccine on the percentage uptake of flu vaccine in the TXT4FLUJAB trial using adjusted CL outcomes, adjusting for individual-level covariates gender, age and presence of disease.

		Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p
No weighting	None	0.133 (−0.016, 0.282)	0.081	0.133 (−0.017, 0.282)	0.082
	HW	(−0.016, 0.282)	0.081	(−0.014, 0.280)	0.077
	SSDF	(−0.019, 0.285)	0.086	(−0.021, 0.286)	0.089
	SSDF + HW	(−0.019, 0.285)	0.086	(−0.018, 0.283)	0.083
Cluster size weighting	None	0.068 (−0.063, 0.198)	0.310	0.071 (−0.058, 0.200)	0.280
	Huber−White	(−0.081, 0.216)	0.372	(−0.069, 0.212)	0.320
	SSDF	(−0.065, 0.201)	0.316	(−0.061, 0.203)	0.288
	SSDF + HW	(−0.084, 0.219)	0.378	(−0.073, 0.215)	0.328
Minimum-variance weighting	None	0.128 (−0.017, 0.273)	0.084	0.128 (−0.017, 0.273)	0.084
	HW	(−0.015, 0.271)	0.080	(−0.014, 0.269)	0.077
	SSDF	(−0.020, 0.275)	0.090	(−0.021, 0.277)	0.091
	SSDF + HW	(−0.018, 0.273)	0.086	(−0.017, 0.273)	0.083

[a]TSLS estimation was adjusted for weekend clinics (yes/no).

Appendix 4. Results for small true LATE

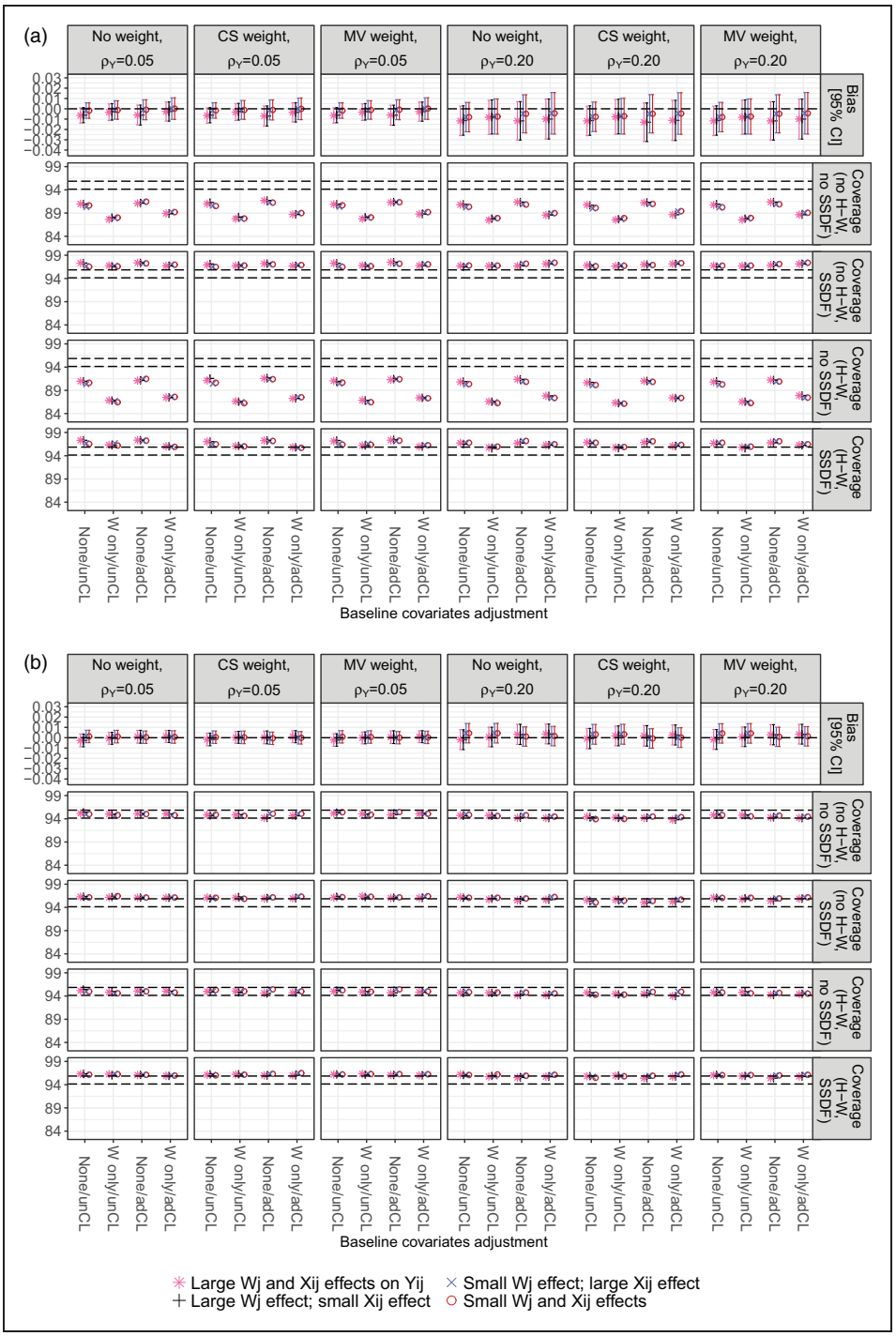


Figure 7. Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with CL non-adherence and small true LATE. Data generation scenarios represented by $*$, $+$, $\times$, and $\circ$. Estimates are obtained via unadjusted or *W*-adjusted TSLS with different weights (none, cluster size (CS), and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for *X* outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panels A and B.

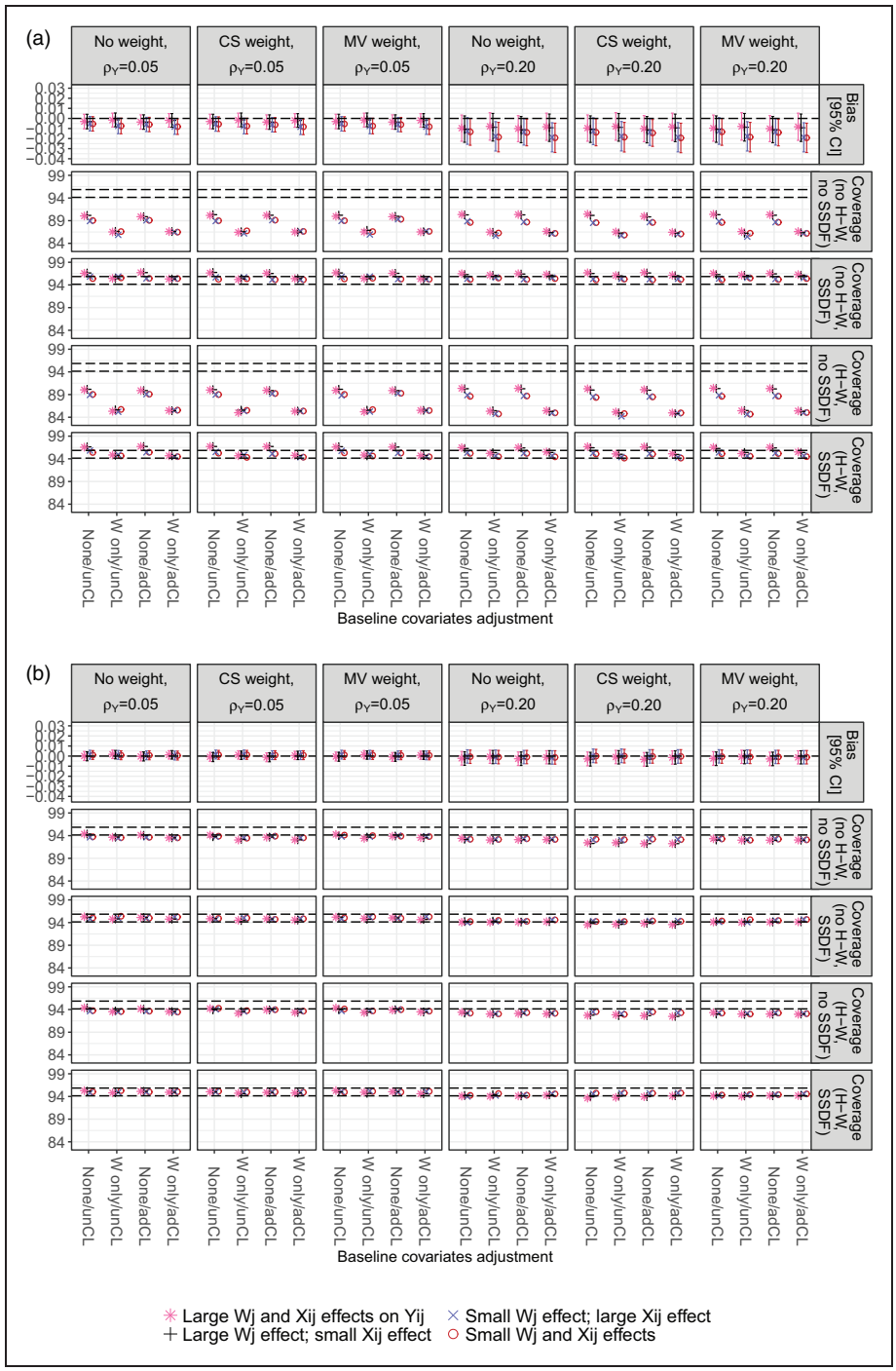


Figure 8. Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and small true LATE. Data generation scenarios represented by $*$, $+$, $\times$, and $\circ$. Estimates are obtained via unadjusted or *W*-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for *X* outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panels A and B.

A.5 Choice of parameters value

To maintain the consistency of the structural model equations used in our simulations, it is necessary to choose a combination or set of parameters values that are suitable. We explain here the approximation strategy used to select the parameters value in our simulations.

1. Adherence at cluster level

Note that R and Z are independent and $\mathbb{E}(R) = \pi = 0.6$, $\text{Var}(R) = \pi(1-\pi) = 0.24$, $\mathbb{E}(Z) = 0.5$ and $\text{Var}(Z) = 0.25$. For simplicity, we chose $\beta_0 = 0$ and $\beta_R = 0$.

The expectation of V_{ij} using the structural model equation is as follows:

$$\begin{aligned}\mathbb{E}(V) &= \beta_{RZ}\mathbb{E}(RZ) + \beta_W\mathbb{E}(W) + \beta_B\mathbb{E}(B) + \mathbb{E}(\upsilon) + \mathbb{E}(\epsilon)\\ &= \beta_{RZ}\mathbb{E}(R)\mathbb{E}(Z) + \beta_W 0 + \beta_B 0 + 0 + 0\\ &= \beta_{RZ}\mathbb{E}(R)\mathbb{E}(Z)\end{aligned} \tag{1}$$

Let σ^2_{res} be the total residuals variance ($\sigma^2_{res} = \sigma^2_\upsilon + \sigma^2_\epsilon$), $\rho_{Y_{con}}$ and $\rho_{Y_{mar}}$ the conditional and marginal ICC for V respectively. $\rho_{Y_{mar}} = \sigma^2_{\upsilon_{mar}}/(\sigma^2_{\upsilon_{mar}} + \sigma^2_{\epsilon_{mar}})$, where $\sigma^2_{\upsilon_{mar}}$ is the unconditional cluster-level variance, $\sigma^2_{\epsilon_{mar}}$ the unconditional individual-level variance.

The variance of V_{ij}, σ^2_V, is given by:

$$\begin{aligned}\sigma^2_V &= \sigma^2_{\upsilon_{mar}} + \sigma^2_{\epsilon_{mar}} = \beta^2_{RZ}\text{Var}(RZ) + \beta^2_W\text{Var}(W) + \beta^2_B\text{Var}(B) + \sigma^2_\upsilon + \sigma^2_\epsilon\\ &= \beta^2_{RZ}\text{Var}(RZ) + \beta^2_W\text{Var}(W) + \beta^2_B\text{Var}(B) + \sigma^2_{res}\\ &= \beta^2_{RZ}\Big(\text{Var}(R)\text{Var}(Z) + \text{Var}(R)[\mathbb{E}(Z)]^2 + \text{Var}(Z)[\mathbb{E}(R)]^2\Big) + \beta^2_W\text{Var}(W)\\ &+ \beta^2_B\rho_B\text{Var}(B) + \beta^2_B(1-\rho_B)\text{Var}(B) + \rho_{V_{con}}\sigma^2_{res} + (1-\rho_{V_{con}})\sigma^2_{res}\end{aligned} \tag{2}$$

$\sigma^2_{\upsilon_{mar}}$ is the sum of cluster-level variances and can be write as follows:

$$\begin{aligned}\sigma^2_{\upsilon_{mar}} = \beta^2_{RZ}\Big(&\text{Var}(R)\text{Var}(Z) + \text{Var}(R)[\mathbb{E}(Z)]^2 + \text{Var}(Z)[\mathbb{E}(R)]^2\Big) +\\ &\beta^2_W\text{Var}(W) + \beta^2_B\rho_B\text{Var}(B) + \rho_{V_{con}}\sigma^2_{res}\end{aligned} \tag{3}$$

When $\sigma^2_V = 1$ and $\beta_{RZ} = \beta_W = \beta_B = \beta$, then $\sigma^2_{\upsilon_{mar}} = \rho_{V_{mar}}$ and equation (2) becomes the equation of an ellipse:

$$\beta^2\Big(\text{Var}(R)\text{Var}(Z) + \text{Var}(R)[\mathbb{E}(Z)]^2 + \text{Var}(Z)[\mathbb{E}(R)]^2 + \text{Var}(W) + \text{Var}(B)\Big) + \sigma^2_{res} = 1 \tag{4}$$

where the major radius a and minor radius b: $b = 1$ and

$$a = \Big[\text{Var}(R)\text{Var}(Z) + \text{Var}(R)[\mathbb{E}(Z)]^2 + \text{Var}(Z)[\mathbb{E}(R)]^2 + \text{Var}(W) + \text{Var}(B)\Big]^{-1/2}.$$

To ensure that $\sigma^2_{res} > 0$ and $\rho_{V_{con}} > 0$ across all scenarios investigated in our simulations, we chose $\beta = 0.1$SD as small effect and $\beta = 0.4$SD as large effect, $\text{Var}(W) = \text{Var}(B) = 0.08$ and $\text{Var}(V) = 1$.

2. Adherence at individual level

Here, $\mathbb{E}(R) = \pi_R = 0.85$, $\text{Var}(R) = \pi_R(1-\pi_R) = 0.1275$. Like in cluster-level adherence setting, $\mathbb{E}(Z) = 0.5$, $\text{Var}(Z) = 0.25$, $\beta_0 = 0$ and $\beta_R = 0$. Equations 1 to 2 remain the same. Because adherence is at individual level, $\text{Var}(R) = \text{Var}(R_{ij}) = \pi_R(1-\pi_R)$ and the variance of the cluster-specific true proportions are given by $\text{Var}(\pi_j) = \rho^{Bin}_R\text{Var}(R)$ as described in Hayes and Moulton

[8], where ρ_R^{Bin} is the ICC for adherence on a probability scale. Therefore, equation 3 becomes:

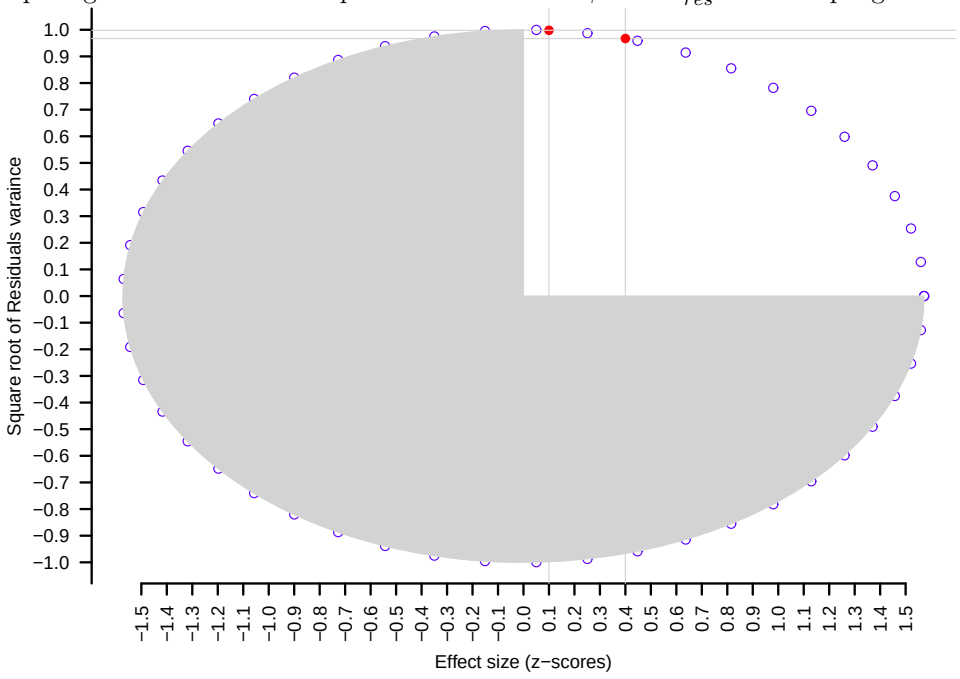
$$\begin{aligned}\sigma^2_{vmar} = \beta^2_{RZ}\Big(\rho_R^{Bin}\text{Var}(R)\text{Var}(Z) + \rho_R^{Bin}\text{Var}(R)[\mathbb{E}(Z)]^2 + \text{Var}(Z)[\mathbb{E}(R)]^2\Big) + \\ \beta^2_W\text{Var}(W) + \beta^2_B\rho_B\text{Var}(B) + \rho_{V_{con}}\sigma^2_{res}\end{aligned} \tag{5}$$

However, we generated R_{ij} using random-intercept logistic regression and set an ICC for adherence in the log-odds scale that we denoted by $\rho_R = 0.50$, implying a level-2 variance of $\pi^2/3$. Then, we approximate the value of ρ_R^{Bin} using the delta method to obtain a between-cluster variance in probability scale, equivalent to the between-cluster variance in log-odds scale as follows:

$$\begin{aligned}\text{Var}[\text{logit}(\pi_{ij})] &\approx \Big(\Big[\log\frac{\pi_R}{1-\pi_R}\Big]'\Big)2\text{Var}(\pi_{ij}) \\ &= \frac{\text{Var}(\pi_{ij})}{[\pi_R(1-\pi_R)]^2} \\ &= \frac{\pi^2}{3}\end{aligned} \tag{6}$$

Therefore, we get $\text{Var}(\pi_{ij}) = \frac{\pi^2}{3}[\pi_R(1-\pi_R)]^2 \approx 0.053$ (lower than $\text{Var}(R)$). Using $\text{Var}(\pi_{ij})$ as the variance of cluster-specific true proportions leads to $\rho_R^{Bin} \approx 0.42$. Equation 4 remains applicable to adherence at individual level setting. Choosing $\beta = 0.1$SD as small effect and $\beta = 0.4$SD as large effect, $\text{Var}(W) = \text{Var}(B) = 0.08$ and $\text{Var}(V) = 1$, also ensure that $\sigma^2_{res} > 0$ and $\rho_{V_{con}} > 0$ for across all scenarios investigated in our simulations.

The ellipse figure below shows the plausible values of of β and σ^2_{res} on the top right.



A.6 Proof of "regression anatomy" formula for OLS estimation

We re-write equation (3.3) referring to $\overline{Y}_j$ simply as Y_j and recall the regression model of Z_j on W_j. We consider W_j to be a single covariate here, for notation simplicity. However, the proof is easily expendable to many covariates. $Y_j = \alpha_0 + \alpha_{\text{Z}}Z_j + \alpha_{\text{W}}W_j + \eta_j$ and $Z_j = \delta_0 + \delta_{\text{W}}W_j + \epsilon_{z_j}$. Then, $\epsilon_{z_j} = Z_j - \delta_0 - \delta_{\text{W}}W_j$. We have by construction, $\text{Cov}(W_j, \eta_j) = 0$, $\text{Cov}(Z_j, \eta_j) = 0$ and

$\text{Cov}(W_j, \epsilon_{z_j}) = 0$.

$$\begin{aligned}\frac{\text{Cov}(Y_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} &= \frac{\text{Cov}(\alpha_0 + \alpha_{\text{Z}} Z_j + \alpha_{\text{W}} W_j + \eta_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} \\ &= \frac{\alpha_{\text{Z}}\text{Cov}(Z_j, \epsilon_{z_j}) + \alpha_{\text{W}}\text{Cov}(W_j, \epsilon_{z_j}) + \text{Cov}(\eta_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} \\ &= \frac{\alpha_{\text{Z}}\text{Cov}(Z_j, \epsilon_{z_j}) + \text{Cov}(\eta_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})}\end{aligned}$$

We have $\text{Cov}(\eta_j, \epsilon_{z_j}) = \text{Cov}(\eta_j, Z_j - \delta_0 - \delta_{\text{W}} W_j) = \text{Cov}(\eta_j, Z_j) - \delta_{\text{W}}\text{Cov}(\eta_j, W_j) = 0$. Then,

$$\frac{\text{Cov}(Y_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} = \frac{\alpha_{\text{Z}}\text{Cov}(Z_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})}$$

We can write $\text{Cov}(Z_j, \epsilon_{z_j}) = \text{Cov}(\delta_0 + \delta_{\text{W}} W_j + \epsilon_{z_j}, \epsilon_{z_j}) = \delta_{\text{W}}\text{Cov}(W_j, \epsilon_{z_j}) + \text{Cov}(\epsilon_{z_j}, \epsilon_{z_j}) = \text{Var}(\epsilon_{z_j})$.

$$\text{Hence, } \frac{\text{Cov}(Y_j, \epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} = \frac{\alpha_{\text{Z}}\text{Var}(\epsilon_{z_j})}{\text{Var}(\epsilon_{z_j})} = \alpha_{\text{Z}}$$

A.7 Proof of "regression anatomy" formula for WLS estimation

The proof is similar to A.6, but with the rescaled variables. We use here equation (3.26) where $\widetilde{\overline{Y}}_j$ is referred to as $\widetilde{Y}_j$ and also the regression model of $\widetilde{Z}_j$ on $\widetilde{W}_j$. Let Ω_j be a variable whose values are the square root of the weight variable ω_j.

$\widetilde{Y}_j = \alpha_0\Omega_j + \alpha_{\text{Z}}\widetilde{Z}_j + \alpha_{\text{W}}\widetilde{W}_j + \widetilde{\eta}_j$ and $\widetilde{Z}_j = \delta_0\Omega_j + \delta_{\text{W}}\widetilde{W}_j + \widetilde{\epsilon}_{z_j}$. Then, $\widetilde{\epsilon}_{z_j} = Z_j - \delta_0\Omega_j - \delta_{\text{W}}\widetilde{W}_j$. By construction, $\text{Cov}(\Omega_j, \widetilde{\eta}_j) = 0$, $\text{Cov}(\widetilde{W}_j, \widetilde{\eta}_j) = 0$, $\text{Cov}(\widetilde{Z}_j, \widetilde{\eta}_j) = 0$, $\text{Cov}(\Omega_j, \widetilde{\epsilon}_{z_j}) = 0$ and $\text{Cov}(\widetilde{W}_j, \widetilde{\epsilon}_{z_j}) = 0$.

$$\begin{aligned}\frac{\text{Cov}(\widetilde{Y}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} &= \frac{\text{Cov}(\alpha_0\Omega_j + \alpha_{\text{Z}}\widetilde{Z}_j + \alpha_{\text{W}}\widetilde{W}_j + \widetilde{\eta}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} \\ &= \frac{\alpha_0\text{Cov}(\widetilde{\Omega}_j, \widetilde{\epsilon}_{z_j}) + \alpha_{\text{Z}}\text{Cov}(\widetilde{Z}_j, \widetilde{\epsilon}_{z_j}) + \alpha_{\text{W}}\text{Cov}(\widetilde{W}_j, \widetilde{\epsilon}_{z_j}) + \text{Cov}(\widetilde{\eta}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} \\ &= \frac{\alpha_{\text{Z}}\text{Cov}(\widetilde{Z}_j, \widetilde{\epsilon}_{z_j}) + \text{Cov}(\widetilde{\eta}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})}\end{aligned}$$

We have $\text{Cov}(\widetilde{\eta}_j, \widetilde{\epsilon}_{z_j}) = \text{Cov}(\widetilde{\eta}_j, \widetilde{Z}_j - \delta_0\Omega_j - \delta_{\text{W}}\widetilde{W}_j) = \text{Cov}(\widetilde{\eta}_j, \widetilde{Z}_j) - \delta_0\text{Cov}(\widetilde{\eta}_j, \Omega_j) - \delta_{\text{W}}\text{Cov}(\widetilde{\eta}_j, \widetilde{W}_j) = 0$. Then,

$$\frac{\text{Cov}(\widetilde{Y}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} = \frac{\alpha_{\text{Z}}\text{Cov}(\widetilde{Z}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})}$$

We have $\text{Cov}(\widetilde{Z}_j, \widetilde{\epsilon}_{z_j}) = \text{Cov}(\delta_0\Omega_j + \delta_{\text{W}}\widetilde{W}_j + \widetilde{\epsilon}_{z_j}, \widetilde{\epsilon}_{z_j}) = \delta_0\text{Cov}(\Omega_j, \widetilde{\epsilon}_{z_j}) + \delta_{\text{W}}\text{Cov}(\widetilde{W}_j, \widetilde{\epsilon}_{z_j}) + \text{Cov}(\widetilde{\epsilon}_{z_j}, \widetilde{\epsilon}_{z_j}) = \text{Var}(\widetilde{\epsilon}_{z_j})$.

$$\text{Hence, } \frac{\text{Cov}(\widetilde{Y}_j, \widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} = \frac{\alpha_{\text{Z}}\text{Var}(\widetilde{\epsilon}_{z_j})}{\text{Var}(\widetilde{\epsilon}_{z_j})} = \alpha_{\text{Z}}$$

A.8 R code for simulated CRT datasets

A.8.1 Generating CRTs with cluster-level adherence

```
### Upload necessary packages
library(foreign) ; library(lme4) ; library(data.table)
library(DataCombine) ; library(foreach)

### SET VALUES FOR MEANS, VARIANCES, ICC AND EFFECTS SIZE
mean.Z<- 0.5; var.W<- 0.08; var.X<- var.W; var.Y<- 1; beta.small<- 0.1
beta.large <- 0.4 ; rho.Ymar.min <- 0.05 ; rho.Ymar.max <- 0.20

### CREATE FUNCTION TO GENERATE CRT
# CRT SIZE PARAMATERS
Nsim     <- NA     # Number of simulations
N.mu     <- NA     # Total number of individual
k        <- NA     # Number of clusters per arm
var.Zj   <- 0.25   # variance of Zj
mean.Zj  <- 0.5 ; k.min <- 5 ; k.max <- 25

# BASELINE COVARIATES PARAMETERS
mu.Wj    <- 0      # mean of level-2 covariate Wj
var.Wj   <- NA     # variance of Wj
mu.Xij   <- 0      # mean of level-1 covariate Xij
rho.Xij  <- 0.05   # ICC of Xij (moderate: 0.05)
var.Xij  <- NA     # Marginal variance of Xij

# ADHERENCE PARAMETERS
pi       <- mean.C # adherence probability for cluster j (0.60)
mean.C   <- 0.6
var.C    <- 0.24   # variance of Cj (0.24)
lambda0  <- log(pi/(1-pi))  # intercept
lambda.W <- NA     # effect of Wj (small: 0.05 ; large: 0.7)
lambda.W.min <- 0.05 ; lambda.W.max <- 0.70

# OUTCOME PARAMETERS (complier=c and never-taker=nt)
beta.0   <- 0   # never-taker clusters mean outcome in control group
beta.C   <- 0   # mean difference c vs. nt in control
beta.Z   <- 0   # randomisation effect in nt (ER assumption)
beta.CZ  <- NA  # randomisation effect in c
beta.W   <- NA  # change in outcome per unit increase in Wj in nt
beta.CW  <- NA  # change in outcome per unit increase in Wj in c
beta.X   <- NA  # change in outcome per unit increase in Xij in nt
beta.CX  <- NA  # change in outcome per unit increase in Xij in c
rho.Ycon <- NA  # marginal ICC (low: 0.05 ; high: 0.20)
var.Yij  <- NA  # Marginal variance of Yij
var.res  <- NA  # Residuals variance

# DIRECTORY PATH PARAMATER
mainDir  <- "H:/"
set      <- ""   # Scenario ID

# CREATE FUNCTION TO SIMULATE CRTs
CRT.1SidCluAdh <-
↪  function(set,N.mu,k,Nsim,lambda.W,beta.CZ,beta.W,beta.CW,beta.X,beta.CX,rho.Ymar,var.Yij,var.Wj,v
↪  {
sim      <- 1  # simulation order
include  <- 0  # Count number of eligible CRT (Dj != Zj)
repeat {
# Set clusters and size
set.seed(sim)
mj   <- rpois(n=1:(2*k), lambda=N.mu/(2*k))
j    <- rep(1:(2*k), times=mj[1:(2*k)])
Zj   <- as.numeric(j>k)
N    <- length(j)       # Total number of individuals in CRT

# Generate level-2 and level-1 explanatory variables
Wj   <- rep(rnorm(2*k, mean=mu.Wj, sd=sqrt(var.Wj))[1:(2*k)], times=mj[1:(2*k)])
Xij  <- rep(rnorm(2*k, mean=mu.Xij, sd=sqrt(rho.Xij*var.Xij))[1:(2*k)],
↪  times=mj[1:(2*k)]) + rnorm(N, mean=mu.Xij, sd=sqrt((1-rho.Xij)*var.Xij))

# Generate true compliance class (compliers and noncompliers)
logit.pi.j <- lambda0 + lambda.W*Wj
pi.j <- 1/(1+exp(-logit.pi.j))
Cj   <- rep(rbinom(2*k, 1, pi.j)[1:(2*k)], times=mj[1:(2*k)])
```

```
# Generate treatment receipt Dj
Dj                <- rep(NA,N)
Dj[Cj==0 & Zj==0] <- Zj[Cj==0 & Zj==0]
Dj[Cj==0 & Zj==1] <- 1-Zj[Cj==0 & Zj==1]
Dj[Cj==1 & Zj==0] <- Zj[Cj==1 & Zj==0]
Dj[Cj==1 & Zj==1] <- Zj[Cj==1 & Zj==1]

# Calculate residuals variance and conditional ICC for Y
var.res  <- var.Yij   - ((var.C*var.Z + var.C*mean.Z^2 + var.Z*mean.C^2)*beta.CZ^2 +
↪  var.W*beta.CW^2 + var.X*beta.CX^2)
rho.Ycon <- (rho.Ymar - ((var.C*var.Z + var.C*mean.Z^2 + var.Z*mean.C^2)*beta.CZ^2 +
↪  var.W*beta.CW^2 + var.X*beta.CX^2*rho.Xij))/var.res

# Generate Yij and Dj, ensure Zj different from Dj, replace CRT otherwise
data.Cj0 <- subset(data.frame(j,k,Cj,Zj,Dj,Wj,Xij), Cj==0)
data.Cj1 <- subset(data.frame(j,k,Cj,Zj,Dj,Wj,Xij), Cj==1)
attach(data.Cj0) ; attach(data.Cj1)
V0       <- rep(1,N)
if (all(Dj==Zj) == "FALSE" & all(Dj==0) == "FALSE" & all(Dj==1) == "FALSE" & all(Cj==V0)
↪  == "FALSE") {
data.Cj0$Yij <- beta.0 + beta.Z*data.Cj0$Zj + beta.W*data.Cj0$Wj + beta.X*data.Cj0$Xij +
↪  rep(rnorm(length(unique(data.Cj0$j)),mean=0, sd=sqrt(rho.Ycon*var.res)),
↪  times=(as.data.frame(table(data.Cj0$j))[,2])[1:length(unique( data.Cj0$j))]) +
↪  rnorm(length(data.Cj0$j), mean=0, sd=sqrt((1-rho.Ycon)*var.res))

data.Cj1$Yij <- beta.C + beta.CZ*data.Cj1$Zj + beta.CW*data.Cj1$Wj +
↪  beta.CX*data.Cj1$Xij +
rep(rnorm(length(unique(data.Cj1$j)),mean=0,sd=sqrt(rho.Ycon*var.res)),
↪  times=(as.data.frame(table(data.Cj1$j))[,2])[1:length(unique( data.Cj1$j))]) +
rnorm(length(data.Cj1$j),mean=0,sd=sqrt((1-rho.Ycon)*var.res))

data <- rbind.data.frame(data.Cj0, data.Cj1)
attach(data)
data <- MoveFront(data, c("j","k","Cj","Zj","Dj","Wj","Xij","Yij"))
subDir1   <- paste0("Scenario", " ", set)
subDir2   <- paste0("Raw")
dir.create(file.path(mainDir, subDir1, subDir2))
include <- include + 1
mydata   <- file.path(mainDir, subDir1, subDir2, paste0("onesided_", set, "_", include,
↪  ".dta"))
write.dta(data, file=mydata) }
sim <- sim + 1
if(include > Nsim - 1) {
break } }
stop.at <- c(sim,include)
return(stop.at) }
```

A.8.2 Generating CRTs with individual-level adherence

```
# SET VALUES FOR MEANS, VARIANCES, ICC AND EFFECTS SIZE
mean.C <- 0.85 ; mean.Z <- 0.5 ; rho.C <- 0.5
var.C <- mean.C*(1-mean.C)   # instead of (pi^2/3)/(1-rho.C)
var.Z <- 0.25 ; var.W <- 0.08 ; var.X <- var.W
var.Y <- 1; beta.small<- 0.1; beta.large<- 0.4
rho.Ymar.min <- 0.05 ; rho.Ymar.max <- 0.20

### CREATE FUNCTION TO GENERATE CRT
# CRT SIZE PARAMATERS
Nsim  <- NA    # Number of simulations
N.mu  <- NA    # Total number of individual on average
k     <- NA    # Number of clusters per arm
k.min <- 5 ; k.max <- 25 ; mean.Zj<- mean.Z
var.Zj<- var.Z # variance of Zj

# BASELINE COVARIATES PARAMETERS
mu.Wj    <- 0    # mean of level-2 covariate Wj
var.Wj   <- NA   # variance of Wj
mu.Xij   <- 0    # mean of level-1 covariate Xij
rho.Xij  <- 0.05 # ICC of Xij (moderate: 0.05)
var.Xij  <- NA   # Marginal variance of Xij

# ADHERENCE BEHAVIOUR PARAMETERS
pij       <- mean.C    # adherence probability
rho.Cij   <- rho.C     # ICC for adherence
var.Cij   <- var.C     # variance of Cj
lambda0   <- log(pij/(1-pij))  # intercept
```

```
lambda.W  <- NA        # effect of Wj (small: 0.05; large: 0.7)
lambda.W.min <- 0.05 ; lambda.W.max <- 0.70
lambda.X  <- NA        # effect of Xij (small: 0.05; large: 0.7)
lambda.X.min <- 0.05 ; lambda.X.max <- 0.70

# OUTCOME PARAMETERS (complier=c; never-taker=nt)
beta.0   <- 0   # nt's mean outcome in control group
beta.C   <- 0   # mean difference c vs. nt outcome in control
beta.Z   <- 0   # randomisation effect in nt (ER assumption)
beta.CZ  <- NA  # randomisation effect in c
beta.W   <- NA  # change in outcome per unit increase in Wj in nt
beta.CW  <- NA  # change in outcome per unit increase in Wj in c
beta.X   <- NA  # change in outcome per unit increase in Xij in nt
beta.CX  <- NA  # change in outcome per unit increase in Xij in c
rho.Ycon <- NA  # marginal ICC (low: 0.05 ; high: 0.20)
var.Yij  <- NA  # Marginal variance of Yij
var.res  <- NA  # Residuals variance

# DIRECTORY PATH PARAMATER
mainDir  <- "H:/"
set      <- ""  # Scenario ID

# CREATE FUNCTION TO SIMULATE CRTs
CRT.1SidIndAdh <- function(set,N.mu,k,Nsim,lambda.W,lambda.X,beta.CZ,beta.W,beta.CW,
↪  beta.X,beta.CX,rho.Ymar,var.Yij,var.Wj,var.Xij,var.Cij,rho.Xij,rho.Cij) {
sim      <- 1   # simulation order
include  <- 0   # Count number of eligible simulated data (Dij different from Zj)
repeat {
# Set clusters and size
set.seed(sim)
mj <- rpois(n=1:(2*k), lambda=N.mu/(2*k))
j  <- rep(1:(2*k), times=mj[1:(2*k)])
Zj <- as.numeric(j>k)
N  <- length(j) # Total number of individuals in CRT

# Generate level-2 and level-1 explanatory variables
Wj  <- rep(rnorm(2*k,mean=mu.Wj,sd=sqrt(var.Wj))[1:(2*k)],times=mj[1:(2*k)])
Xij <- rep(rnorm(2*k, mean=mu.Xij, sd=sqrt(rho.Xij*var.Xij))[1:(2*k)],
↪  times=mj[1:(2*k)])+rnorm(N,mean=mu.Xij,sd=sqrt((1-rho.Xij)*var.Xij))

# Generate true compliance class (compliers and noncompliers)
zeta.j <- rep(rnorm(2*k,mean=0,sd=sqrt((rho.Cij*var.Cij))),times=mj[1:(2*k)])
logit.pi.ij  <- lambda0 + lambda.W*Wj + lambda.X*Xij + zeta.j
pi.ij  <- 1/(1+exp(-logit.pi.ij))
Cij    <- rbinom(N, 1, pi.ij)

# Generate treatment receipt Dij (one-sided non-adherence)
Dij <- rep(NA,N)
Dij[Cij==0 & Zj==0] <- Zj[Cij==0 & Zj==0]
Dij[Cij==0 & Zj==1] <- 1-Zj[Cij==0 & Zj==1]
Dij[Cij==1 & Zj==0] <- Zj[Cij==1 & Zj==0]
Dij[Cij==1 & Zj==1] <- Zj[Cij==1 & Zj==1]

# Calculate residuals variance and conditional ICC for Y
var.res  <- var.Yij   - ((var.Cij*var.Z + var.Cij*mean.Z^2 + var.Z*mean.C^2)*beta.CZ^2 +
↪  var.Wj*beta.CW^2 + var.Xij*beta.CX^2)
rho.Ycon <- (rho.Ymar - ((rho.Cij*var.Cij*var.Z + rho.Cij*var.Cij*mean.Z^2 +
↪  var.Z*mean.C^2)*beta.CZ^2 + var.Wj*beta.CW^2 + var.Xij*beta.CX^2*rho.Xij))/var.res

# Generate outcome Yij and treatment receipt and ensure Zj different from Dij, replace
↪  datat set otherwise
data.Cij0 <- subset(data.frame(j,k,Cij,Zj,Dij,Wj,Xij), Cij==0)
data.Cij1 <- subset(data.frame(j,k,Cij,Zj,Dij,Wj,Xij), Cij==1)
attach(data.Cij0) ; attach(data.Cij1)

V0 <- rep(1,N)
if (all(Dij==Zj) == "FALSE" & all(Dij==0) == "FALSE" & all(Dij==1) == "FALSE" &
↪  all(Cij==V0) == "FALSE") {
data.Cij0$Yij <- beta.0 + beta.Z*data.Cij0$Zj + beta.W*data.Cij0$Wj + beta.X*
↪  data.Cij0$Xij+rep(rnorm(length(unique(data.Cij0$j)),mean=0,
↪  sd=sqrt(rho.Ycon*var.res)),
↪  times=(as.data.frame(table(data.Cij0$j))[,2])[1:length(unique( data.Cij0$j))]) +
rnorm(length(data.Cij0$j),mean=0,sd=sqrt((1-rho.Ycon)*var.res))

data.Cij1$Yij <- beta.C + beta.CZ*data.Cij1$Zj + beta.CW*data.Cij1$Wj +
↪  beta.CX*data.Cij1$Xij +
```

```
rep(rnorm(length(unique(data.Cij1$j)),mean=0, sd=sqrt(rho.Ycon*var.res)),
↪  times=(as.data.frame(table(data.Cij1$j))[,2])[1:length( unique(data.Cij1$j))]) +
rnorm(length(data.Cij1$j), mean=0, sd=sqrt((1-rho.Ycon)*var.res))

data <- rbind.data.frame(data.Cij0, data.Cij1)
attach(data)
data <- MoveFront(data, c("j","k","Cij","Zj","Dij","Wj","Xij","Yij"))
subDir1   <- paste0("Scenario", " ", set)
subDir2   <- paste0("Raw")
dir.create(file.path(mainDir, subDir1, subDir2))
include   <- include + 1
mydata    <- file.path(mainDir, subDir1, subDir2, paste0("onesided_", set, "_", include,
↪  ".dta"))
write.dta(data, file=mydata) }
sim <- sim + 1
if(include > Nsim - 1) {
break } }
stop.at <- c(sim,include)
return(stop.at) }
```

A.9 Stata code for CL-TSLS and Schochet-Chiang method

Variables definition
i: Individual unit ID ; j: Cluster unit ID ; Y_{ij}: Outcome ; D_{ij}: Treatment received ;
W_j: Baseline cluster-level covariate and X_{ij}: Baseline individual-level covariate

A.9.1 CL-TSLS with covariate adjustment, using unadjusted CL summaries

```
* Estimate marginal ICC for Yij
qui mixed Yij || j:, reml
scalar rhoY_noEs = exp(2*[lns1_1_1]_cons)/(exp(2*[lns1_1_1]_cons) +
↪  exp(2*[lnsig_e]_cons))

* Generate cluster summaries
gen clusterid = j
collapse (count) nobs=j (mean) Zj=Zj (mean) propDij=Dij (mean) meanWj=Wj (mean)
↪  meanYij=Yij, by(clusterid)

* Create minimum variance weights
gen mvw_noE = nobs/(1+rhoY_noEs*(nobs-1))

* No weighting, with Huber-White SEs and SSDF adjustment
ivregress 2sls meanYij meanWj (propDij = Zj), robust small

* Cluster size weighting, with Huber-White SEs and SSDF adjustment
ivregress 2sls meanYij meanWj (propDij = Zj) [aw=nobs], robust small

* Minimum-variance weighting, with Huber-White SEs and SSDF adjustment
ivregress 2sls meanYij meanWj (propDij = Zj) [aw=mvw_noE], robust small
```

A.9.2 CL-TSLS with covariate adjustment, using adjusted CL summaries

```
* Generate predicted level-1 residuals adjusted for Xij, using OLS
qui regress        Yij Xij
predict        Ehatij, r

* Estimate marginal ICC for Ehatij
qui mixed Ehatij || j:, reml
scalar rhoEhat_noEs = exp(2*[lns1_1_1]_cons)/(exp(2*[lns1_1_1]_cons) +
↪  exp(2*[lnsig_e]_cons))

* Generate CL summaries
gen clusterid = j
collapse (count) nobs=j (mean) Zj=Zj (mean) propDij=Dij (mean) meanWj=Wj (mean)
↪  meanEhatij=Ehatij, by(clusterid)

* Create minimum variance weights
gen mvw_noE = nobs/(1+rhoEhat_noEs*(nobs-1))

* No weighting, with Huber-White SEs and SSDF adjustment
```

```
ivregress 2sls meanEhatij meanWj (propDij = propZj), robust small

* Cluster size weighting, with Huber-White SEs and SSDF adjustment
ivregress 2sls meanEhatij meanWj (propDij = propZj) [aw=nobs], robust small

* Minimum-variance weighting, with Huber-White SEs and SSDF adjustment
ivregress 2sls meanEhatij meanWj (propDij = propZj) [aw=mvw_noE], robust small
```

A.9.3 Schochet-Chiang method

```
*** Macro for estimating CL-LATE without covariate adjustment
preserve
collapse (mean) Yij (mean) Dij, by(j Zj)
regress Yij i.Zj
matrix B_ITTy_noE = e(b)
matrix V_ITTy_noE = e(V)
regress Dij i.Zj
matrix B_ITTd_noE = e(b)
matrix V_ITTd_noE = e(V)
scalar cace = el(B_ITTy_noE,1,2)/el(B_ITTd_noE,1,2)

* Estimate variance of CACE
/* Predict cluster-level outcome and treatment receipt */
qui regress        Yij i.Zj
scalar k = e(rank)
predict Yij_hat if e(sample), xb
qui regress        Dij i.Zj if Yij<.
predict        Dij_hat if e(sample), xb

/* Store number of clusters in control and in active group (only account for clusters
↪  included in analysis) */
qui regress        Yij i.Zj if Zj==0
scalar        K_control = e(N)
qui regress        Yij i.Zj if Zj==1
scalar        K_active  = e(N)

/* Calculate variance of ITTd */
gen        rsq = (Dij-Dij_hat)^2
total rsq if Zj==0
scalar var_ITTd_control= el(r(table),1,1)/((K_control-k)*K_control)
total rsq if Zj==1
scalar var_ITTd_active = el(r(table),1,1)/((K_active -k)*K_active )
scalar var_ITTd = var_ITTd_control + var_ITTd_active
drop rsq

/* Calculate covariance of ITTy and ITTd */
gen        ydsq = (Yij-Yij_hat)*(Dij-Dij_hat)
total ydsq if Zj==0
scalar covar_ITTyd_control= el(r(table),1,1)/((K_control-k)*K_control)
total ydsq if Zj==1
scalar covar_ITTyd_active = el(r(table),1,1)/((K_active -k)*K_active )
scalar covar_ITTyd = covar_ITTyd_control + covar_ITTyd_active
drop ydsq

/* Calculate variance of CACE */
scalar var_cace = el(B_ITTd_noE,1,2)^(-2) * (el(V_ITTy_noE,2,2) + cace^2*var_ITTd -
↪  2*cace*covar_ITTyd)

/* Store results */
scalar se = sqrt(var_cace)
scalar ll95ci = cace - invnormal(0.975)*se
scalar ul95ci = cace + invnormal(0.975)*se

*** Macro for estimating CL-LATE with covariate adjustment
preserve
collapse (mean) Yij (mean) Dij (mean) Wj, by(j Zj)
regress        Yij i.Zj Wj
matrix B_ITTy = e(b)
matrix V_ITTy = e(V)
regress Dij i.Zj Wj
matrix B_ITTd = e(b)
matrix V_ITTd = e(V)
scalar cace = el(B_ITTy,1,2)/el(B_ITTd,1,2)

* Estimate variance of CACE
/* Predict cluster-level outcome and treatment receipt */
```

```
qui regress        Yij i.Zj Wj
scalar        k = e(rank)
predict        Yij_hat if e(sample), xb
qui regress        Dij i.Zj Wj         if Yij<.
predict        Dij_hat if e(sample), xb

/* Store number of clusters in control and active groups */
qui regress        Yij i.Zj Wj        if Zj==0
scalar K_control = e(N)
qui regress        Yij i.Zj Wj        if Zj==1
scalar K_active  = e(N)

/* Calculate variance of ITTd */
gen        rsq = (Dij-Dij_hat)^2
total rsq if Zj==0
scalar var_ITTd_control= el(r(table),1,1)/((K_control-k)*K_control)
total rsq if Zj==1
scalar var_ITTd_active = el(r(table),1,1)/((K_active -k)*K_active )
scalar var_ITTd = var_ITTd_control + var_ITTd_active
drop rsq

/* Calculate covariance of ITTy and ITTd */
gen        ydsq = (Yij-Yij_hat)*(Dij-Dij_hat)
total ydsq if Zj==0
scalar covar_ITTyd_control= el(r(table),1,1)/((K_control-k)*K_control)
total ydsq if Zj==1
scalar covar_ITTyd_active = el(r(table),1,1)/((K_active -k)*K_active )
scalar covar_ITTyd = covar_ITTyd_control + covar_ITTyd_active
drop ydsq

/* Calculate variance of CACE */
scalar var_cace = el(B_ITTd,1,2)^(-2) * (el(V_ITTy,2,2) + cace^2*var_ITTd -
↪  2*cace*covar_ITTyd)

/* Store results */
scalar se = sqrt(var_cace)
scalar ll95ci = cace - invnormal(0.975)*se
scalar ul95ci = cace + invnormal(0.975)*se
```

A.10 Code for TSLS, Wald and Bayesian estimations

A.10.1 Wald estimation with covariate adjustment for cluster-level adherence

```
*** Macro for estimating IL-LATE
cap program drop illateest
program def illateest, eclass
marksample touse

/* Estimate of ITT effect on outcome as risk difference */
qui reg        Yij i.Zj Wj Xij
scalar ITTy = _b[1.Zj]

/* Estimate of ITT effect on treatment received as risk difference */
preserve
egen pickonecluster = tag(j)
qui reg        Dj i.Zj Wj         if pickonecluster==1
scalar ITTd = _b[1.Zj]
restore

/* IL-LATE point estimate */
mat        late = ITTy/ITTd
mat        colnames late = "late"
eret post late, e(`touse')
eret local cmd illateest
eret display
end

*** Run Macro for estimating IL-LATE with bootstrap
use "...\dataset.dta", clear
gen id = _n
```

```
bootstrap _b[late], cluster(j) idcluster(j_newid) group(id) strata(Z) reps(1500)
↪  seed(123456789): illateest
mat CI = e(ci_normal)
mat B = e(b)

/* Store results */
scalar LATE = B[1,1]
scalar ll95ci = CI[1,1]
scalar ul95ci = CI[2,1]
```

A.10.2 Wald estimation with covariate adjustment for individual-level adherence

```
*** Macro for estimating IL-LATE
cap program drop illateest
program def illateest, rclass

/* Estimate of ITT effect on outcome as risk difference */
qui reg        Yij i.Zj Wj Xij
matrix B_ITTy = e(b)
local ITTy = el(B_ITTy,1,2)

/* Estimate of ITT effect on treatment received as risk difference */
qui reg        Dij i.Zj Wj Xij
matrix B_ITTd = e(b)
local ITTd = el(B_ITTd,1,2)

/* IL-LATE point estimate */
return scalar late = `ITTy'/`ITTd'
end

*** Run Macro for estimating IL-LATE with bootstrap
use "...\dataset.dta", clear
gen id = _n
bootstrap r(late), cluster(j) idcluster(j_newid) group(id) strata(Z) reps(1500)
↪  seed(123456789): illateest
mat CI = e(ci_normal)
mat B = e(b)

/* Store results */
scalar LATE = B[1,1]
scalar ll95ci = CI[1,1]
scalar ul95ci = CI[2,1]
```

A.10.3 TSLS with HWR SEs and covariate adjustment

```
*** Estimate IL-LATE
use "...\dataset.dta", clear
ivregress 2sls Yij Wj Xij (Dij = i.Zj), vce(cluster j)
matrix B_HWR = e(b)
matrix V_HWR = e(V)

/* Store results */
scalar LATE = B_HWR[1,1]
scalar ll95ci = B_HWR[1,1] - invnorm(0.975)*sqrt(V_HWR[1,1])
scalar ul95ci = B_HWR[1,1] + invnorm(0.975)*sqrt(V_HWR[1,1])
```

A.10.4 TSLS with Moulton's SEs and covariate adjustment

```
*** Estimate IL-LATE
use ".../dataset.dta", clear

/* predict Y residuals and estimate its ICC */
ivregress 2sls Yij Wj Xij (Dij = i.Zj), first
predict        Yres, resid
loneway        Yres j
scalar rho_Yres = r(rho)

/* Fit 1st stage regression, predict Dhat and estimate ICC for Dhat */
regress        Dij i.Zj Wj Xij
predict        Dhat2, xb
loneway        Dhat j
scalar rho_Dhat = r(rho)

/* Estimate Var(cluster size) and Mean(cluster size) */
```

```
bysort j: egen n_j = count(Zj)
egen pickonecluster = tag(j)
sum        n_j if pickonecluster, d
scalar Var_n_j  = r(Var)
scalar Mean_n_j = r(mean)

/* Calculate Moulton factor */
scalar Moulton_factor = sqrt(1 + (((Var_n_j/Mean_n_j) + Mean_n_j -
↪  1)*rho_Dhat*rho_Yres))

/* Obtain SEs from conventional TSLS i.e. ignoring clustering */
ivregress 2sls Yij Wj Xij (Dij = i.Zj)
matrix B_Moulton = e(b)
matrix V = e(V)

/* Store results */
scalar LATE = B_Moulton[1,1]
scalar ll95ci = B_Moulton[1,1] - invnorm(0.975)*sqrt(V[1,1] * Moulton_factor^2)
scalar ul95ci = B_Moulton[1,1] + invnorm(0.975)*sqrt(V[1,1] * Moulton_factor^2)
```

A.10.5 Bayesian multilevel mixture model with covariate adjustment

```
library(foreign) ; library(base) ; library(rjags)
library(coda) ; library(mcmcplots)

data    <- read.dta(".../dataset.dta")
data$Rij<- data$Dij
data$Rij[data$Zj==0] <- NA
Nlevel1 <- nrow(data)
Nlevel2 <- length(unique(data$j,incomparable=F))

# Set burn-in and iterations
n.chains<- 2; n.iter<- 50000; n.burnin<- 5000; n.thin <- 1
write("model {for (i in 1:Nlevel1) {

# Level-1 outcome model specification
Yij[i]  ~  dnorm(mu_Yij[i,Cij[i]],tau_Yij)
mu_Yij[i,1]   <- B0[1] + b1[j[i]] + B.x[1]*Xij[i] + B.w[1]*Wj[i]
mu_Yij[i,2]   <- B0[2] + b2[j[i]] + B2[2]*Zj[i] + B.x[2]*Xij[i] + B.w[2]*Wj[i]

# Level-1 compliance class model specification
Rij[i]  ~  dbern(pi[i])
pi[i]   <- ilogit(A[i])
A[i]    <- A0 + a[j[i]] + A.x*Xij[i] + A.w*Wj[i]
Cij[i]  ~  dcat(p[i,])
p[i,2]  <- pi[i]
p[i,1]  <- 1-pi[i]
}

for (j in 1:Nlevel2) {
# Level-2 random-effects distribution
b1[j]   ~  dnorm(0,tau_b)
b2[j]   ~  dnorm(0,tau_b)
a[j]    ~  dnorm(0,tau_a)
}

# Priors specification for random effects
tau_b   <- 1/pow(sigma_b,2)
sigma_b ~  dt(0,pow(100,-2),1)T(0,)
tau_a   <- 1/pow(sigma_a,2)
sigma_a ~  dt(0,pow(100,-2),1)T(0,)

# Priors specification for outcome distribution
sigma_Yij   <- 1/sqrt(tau_Yij)
tau_Yij     ~  dgamma(0.001,0.001)

# Priors specification for outcome model
A0 ~ dnorm(0,0.001); B0[1] ~ dnorm(0,0.001)
B0[2] ~ dnorm(0,0.001); B2[2] ~ dnorm(0,0.001)
A.x ~ dnorm(0,0.001); B.x[1] ~ dnorm(0,0.001)
B.x[2] ~ dnorm(0,0.001); A.w ~ dnorm(0,0.001)
B.w[1] ~ dnorm(0,0.001); B.w[2] ~ dnorm(0,0.001)
}", "simModel.jags")

rjags.model <- jags.model(simModel.jags", data=list(Nlevel1=Nlevel1, Nlevel2=Nlevel2, j=data$j, Zj=da
rjags.par <- c("A0","B0[1]","B0[2]","B2[2]", "sigma_a", "sigma_b", "sigma_Yij")
rjags.sim <- coda.samples(rjags.model, rjags.par, n.burn=n.burnin, n.iter=n.iter, thin=n.thin)
result <- summary(rjags.sim)
ess    <- effectiveSize(rjags.sim)
gelman <- gelman.diag(rjags.sim)

# Store results
LATE <- result[["quantiles"]][4,3]
LATEperc2.5 <- result[["quantiles"]][4,1]
LATEperc97.5 <- result[["quantiles"]][4,5]
ess <- ess[4]
gelman <- gelman[["psrf"]][4]
```

A.11 Two-level multiple imputation codes using the "jomo" package in R

Multilevel joint modelling multiple imputation [122,123] was used to handle missing data, assuming missingness at random *i.e.* the probability of observing data is the same for all participants

conditional on observed covariates and outcome variable. The multilevel multiple imputation was carried out using "jomo" package in R [124] and done separately for the control and the active groups. We generated a large number of imputations (250 imputed data sets) to minimise the Monte Carlo error. We discarded the 700 000 first cycles and chose 1 000 iterations between two successive imputations. Imputed data were exported to Stata 15 for analyses. Estimates were pooled using the Rubin's rule [126].

```
library(foreign) ; library(jomo)
data <- read.dta(".../Opera.dta")
data$RECEIVED<- as.factor(data$D)
data$HOME1 <- as.numeric(data$HOME=="Private & Nursing")
data$HOME2 <- as.numeric(data$HOME=="Voluntary/LA")
data.control <- subset(data, ALLOC=="Control")
data.active <- subset(data, ALLOC=="Intervention")
Nburnin <- 700000 ; Niterbtwn <- 1000 ; M <- 250
```

A. Convergence and imputation in control group only

```
Y1 <- data.frame(data.control$SPPB2, data.control$AGE,
data.control$SPPB0, data.control$MMSE0, data.control$ANTIDEP0)
X1 <- data.frame(data.control$RECEIVED, data.control$SEX,
data.control$PLACE, data.control$SIZE, data.control$HOME1,
data.control$HOME2)
clusterid <- data.control$ HOME_ID
```

A.1. Convergence check in control group

```
imp.chk.control <- jomo.MCMCchain(Y=Y1, X=X1, clus=clusterid,
nburn=Nburnin, meth="common")
imp.chk.control$collectbeta[,,1]
imp.chk.control$collectcovu[,,1]
```

A.2. Run imputation model in control group

```
imp.model.control <- jomo(Y=Y1, X=X1, clus=clusterid, nburn=Nburnin,
nbetween=Niterbtwn, nimp=M, meth="common")
imp.model.control$Z <- 0
names(imp.model.control) <- gsub("data.control.", "",
names(imp.model.control))
```

B. Convergence and imputation in active group only

```
Y1 <- data.frame(data.active$SPPB2, data.active$AGE,
data.active$SPPB0, data.active$MMSE0, data.active$ANTIDEP0)
X1 <- data.frame(data.active$RECEIVED, data.active$SEX,
data.active$PLACE, data.active$SIZE, data.active$HOME1,
data.active$HOME2)
clusterid <- data.active$ HOME_ID
```

B.1. Convergence check in active group

```
imp.chk.active <- jomo.MCMCchain(Y=Y1, X=X1, clus=clusterid,
nburn=Nburnin, meth="common")
imp.chk.active$collectbeta[,,1]
imp.chk.active$collectcovu[,,1]
```

B.2. Run imputation model in active group

```
imp.model.active <- jomo(Y=Y1, X=X1, clus=clusterid, nburn=Nburnin,
nbetween=Niterbtwn, nimp=M, meth="common")
imp.model.active$Z <- 1
names(imp.model.active) <- gsub("data.active.", "",
names(imp.model.active))
```

C. Combine control and active groups imputed data and export data set to Stata

```
data.imp <- rbind(imp.model.control,imp.model.active)
write.dta(data.imp, ``.../Opera_jomo.dta")
```

A.12 Sensitivity analyses code for the OPERA trial

A.12.1 TSLS with Huber-White-Rogers and Moulton's SEs

```
use        ".../Opera.dta", clear

*** Estimate ITT effect
mixed sppb2 i.Z i.place i.size i.hometype mmse0 agebaseline i.sex sppb0 i.antidep0 ||
↪  pat_hmeid:, reml
matrix B_ITTy_E = e(b)

*** Set lambda to be 5% of ITT effect estimate
scalar lambda = 0.05*el(B_ITTy_noE,1,2)

*** Generate adjusted outcome (sppb2_adj)
gen sppb2_adj = sppb2
replace sppb2_adj = sppb2 + `lambda' *D1

*** TSLS with Huber-White-Rogers SEs
ivregress 2sls sppb2_adj i.place i.size i.hometype mmse0 agebaseline i.sex sppb0
↪  i.antidep0 (D1 = i.Z), vce(cluster pat_hmeid)

*** TSLS with Moulton' SEs
/* predict Y residuals and estimate its ICC */
ivregress 2sls sppb2_adj i.place i.size i.hometype mmse0 agebaseline i.sex sppb0
↪  i.antidep0 (D1 = i.Z), first
predict        Yres, resid
loneway        Yres pat_hmeid
scalar        rho_Yres = r(rho)

/* Fit 1st stage regression, predict Dhat and estimate ICC for Dhat */
regress D1 i.Z i.place i.size i.hometype mmse0 agebaseline i.sex sppb0 i.antidep0
predict        Dhat, xb
loneway        Dhat pat_hmeid
scalar        rho_Dhat = r(rho)

/* Estimate Var(cluster size) nd Mean(cluster size) */
bysort pat_hmeid: egen n_j = count(pat_id)
egen pickonecluster = tag(pat_hmeid)
sum        n_j if pickonecluster, d
scalar Var_n_j  = r(Var)
scalar Mean_n_j = r(mean)

/* Calculate Moulton factor */
scalar Moulton_factor = sqrt(1 + (((Var_n_j/Mean_n_j) + Mean_n_j -
↪  1)*rho_Dhat*rho_Yres))

/* Obtain SEs from conventional TSLS i.e. ignoring clustering */
ivregress 2sls sppb2_adj i.place i.size i.hometype mmse0 agebaseline i.sex sppb0
↪  i.antidep0 (D1 = i.Z)
matrix B = e(b)
matrix V = e(V)

/* Store results */
scalar LATE = B[1,1]
scalar ll95ci = B[1,1] - invnorm(0.975)*sqrt(V[1,1] * Moulton_factor^2)
scalar ul95ci = B[1,1] + invnorm(0.975)*sqrt(V[1,1] * Moulton_factor^2)
```

A.12.2 Bayesian multilevel mixture with local-to-0 prior

```
library(foreign); library(base); library(rjags); library(coda)
library(mcmcplots)

data          <- read.dta(".../Opera.dta")
data$Z        <- NULL
data$Z        <- as.numeric(data$alloc=="Intervention")
data$R1       <- data$D1
data$R1[data$alloc=="Control"] <- NA
Nlevel1       <- nrow(data)
Nlevel2       <- length(unique(data$pat_hmeid,incomparable=F))
data$home_id  <- as.numeric(as.factor(data$pat_hmeid))
data$age      <- data$agebaseline
data$antidep0 <- as.numeric(data$antidep0=="Y")
```

```
data$place    <- as.numeric(data$place=="London")
data$size     <- as.numeric(data$size==">=32 beds")
data$hometype1<- as.numeric(data$hometype=="Private & Nursing")
data$hometype2<- as.numeric(data$hometype=="Voluntary/LA")
data$sex      <- as.numeric(data$sex=="M")

### Set burn-in and iterations
n.chains<- 3; n.iter<- 900000; n.burnin<- 100000; n.thin<- 1

### 1. Assuming no ER and variance homogeneity across latent classes
# 1.1. Adjustment for baseline covariates with missing values and use of level-2 Sigma
↪  half-Cauchy prior
write("
model {
for (i in 1:Nlevel1) {

# Level-1 missing covariates model specification in increasing order of missingness
antidep0[i]~  dbern(pi_antidep0[i,C[i]])
pi_antidep0[i,1]    <- ilogit(m1[1] + m1_b1[home_id[i]] + m1_2[1]*Z[i] + m1_3[1]*sex[i]
↪  + m1_4[1]*place[i] + m1_5[1]*size[i] + m1_6[1]*hometype1[i] + m1_7[1]*hometype2[i])
pi_antidep0[i,2]    <- ilogit(m1[2] + m1_b2[home_id[i]] + m1_2[2]*Z[i] + m1_3[2]*sex[i]
↪  + m1_4[2]*place[i] + m1_5[2]*size[i] + m1_6[2]*hometype1[i] + m1_7[2]*hometype2[i])

age[i]     ~  dnorm(mu_age[i,C[i]],tau_age)
mu_age[i,1]<- m2[1] + m2_b1[home_id[i]] + m2_2[1]*Z[i] + m2_3[1]*sex[i] +
↪  m2_4[1]*place[i] + m2_5[1]*size[i] + m2_6[1]*hometype1[i] + m2_7[1]*hometype2[i] +
↪  m2_8[1]*antidep0[i]
mu_age[i,2]<- m2[2] + m2_b2[home_id[i]] + m2_2[2]*Z[i] + m2_3[2]*sex[i] +
↪  m2_4[2]*place[i] + m2_5[2]*size[i] + m2_6[2]*hometype1[i] + m2_7[2]*hometype2[i] +
↪  m2_8[2]*antidep0[i]

sppb0[i]   ~  dnorm(mu_sppb0[i,C[i]],tau_sppb0)
mu_sppb0[i,1]       <- m3[1] + m3_b1[home_id[i]] + m3_2[1]*Z[i] + m3_3[1]*sex[i] +
↪  m3_4[1]*place[i] + m3_5[1]*size[i] + m3_6[1]*hometype1[i] + m3_7[1]*hometype2[i] +
↪  m3_8[1]*antidep0[i] + m3_9[1]*age[i]
mu_sppb0[i,2]       <- m3[2] + m3_b2[home_id[i]] + m3_2[2]*Z[i] + m3_3[2]*sex[i] +
↪  m3_4[2]*place[i] + m3_5[2]*size[i] + m3_6[2]*hometype1[i] + m3_7[2]*hometype2[i] +
↪  m3_8[2]*antidep0[i] + m3_9[2]*age[i]

mmse0[i]   ~  dnorm(mu_mmse0[i,C[i]],tau_mmse0)
mu_mmse0[i,1]       <- m4[1] + m4_b1[home_id[i]] + m4_2[1]*Z[i] + m4_3[1]*sex[i] +
↪  m4_4[1]*place[i] + m4_5[1]*size[i] + m4_6[1]*hometype1[i] + m4_7[1]*hometype2[i] +
↪  m4_8[1]*antidep0[i] + m4_9[1]*age[i] + m4_10[2]*sppb0[i]
mu_mmse0[i,2]       <- m4[2] + m4_b2[home_id[i]] + m4_2[2]*Z[i] + m4_3[2]*sex[i] +
↪  m4_4[2]*place[i] + m4_5[2]*size[i] + m4_6[2]*hometype1[i] + m4_7[2]*hometype2[i] +
↪  m4_8[2]*antidep0[i] + m4_9[2]*age[i] + m4_10[2]*sppb0[i]

# Level-1 outcome model specification
sppb2[i]   ~  dnorm(mu_sppb2[i,C[i]],tau_sppb2)
mu_sppb2[i,1] <- B0[1] + b1[home_id[i]] + B2[1]*Z[i] + B3[1]*sex[i] + B4[1]*place[i] +
↪  B5[1]*size[i] + B6[1]*hometype1[i] + B7[1]*hometype2[i] + B8[1]*antidep0[i] +
↪  B9[1]*age[i] + B10[1]*sppb0[i] + B11[1]*mmse0[i]
mu_sppb2[i,2] <- B0[2] + b2[home_id[i]] + B2[2]*Z[i] + B3[2]*sex[i] + B4[2]*place[i] +
↪  B5[2]*size[i] + B6[2]*hometype1[i] + B7[2]*hometype2[i] + B8[2]*antidep0[i] +
↪  B9[2]*age[i] + B10[2]*sppb0[i] + B11[2]*mmse0[i]

# Level-1 compliance class model specification
R[i]  ~  dbern(pi[i])
pi[i] <- ilogit(A[i] + A3*sex[i] + A4*place[i] + A5*size[i] + A6*hometype1[i] +
↪  A7*hometype2[i])
A[i]  <- A0 + a[home_id[i]]
C[i]  ~  dcat(p[i,]); p[i,2]<- pi[i]; p[i,1]<- 1-pi[i]
}

for (j in 1:Nlevel2) {
# Level-2 random-effects distribution
b1[j] ~ dnorm(0,tau_b); b2[j] ~ dnorm(0,tau_b)
a[j] ~  dnorm(0,tau_a); m1_b1[j] ~ dnorm(0,tau_m1_b)
m1_b2[j] ~ dnorm(0,tau_m1_b); m2_b1[j] ~ dnorm(0,tau_m2_b)
m2_b2[j] ~ dnorm(0,tau_m2_b); m3_b1[j] ~ dnorm(0,tau_m3_b)
m3_b2[j] ~ dnorm(0,tau_m3_b); m4_b1[j] ~ dnorm(0,tau_m4_b)
m4_b2[j] ~ dnorm(0,tau_m4_b)
}

# Priors specification for random effects
tau_b <- 1/pow(sigma_b,2); sigma_b ~ dt(0,pow(100,-2),1)T(0,)
```

```
tau_a <- 1/pow(sigma_a,2); sigma_a ~ dt(0,pow(100,-2),1)T(0,)
tau_m1_b <- 1/pow(sigma_m1_b,2); sigma_m1_b ~ dt(0,pow(100,-2),1)T(0,)
tau_m2_b <- 1/pow(sigma_m2_b,2); sigma_m2_b ~ dt(0,pow(100,-2),1)T(0,)
tau_m3_b <- 1/pow(sigma_m3_b,2); sigma_m3_b ~ dt(0,pow(100,-2),1)T(0,)
tau_m4_b <- 1/pow(sigma_m4_b,2); sigma_m4_b ~ dt(0,pow(100,-2),1)T(0,)

# Priors specification for outcome distribution
sigma_sppb2 <- 1/sqrt(tau_sppb2); tau_sppb2 ~ dgamma(0.001,0.001)
sigma_age <- 1/sqrt(tau_age); tau_age ~ dgamma(0.001,0.001)
sigma_sppb0 <- 1/sqrt(tau_sppb0); tau_sppb0 ~ dgamma(0.001,0.001)
sigma_mmse0 <- 1/sqrt(tau_mmse0); tau_mmse0 ~ dgamma(0.001,0.001)

# Priors specification for outcome model
A0 ~ dnorm(0,0.001); A3 ~ dnorm(0,0.001); A4 ~ dnorm(0,0.001)
A5 ~ dnorm(0,0.001); A6 ~ dnorm(0,0.001); A7 ~ dnorm(0,0.001)
B0[1] ~ dnorm(0,0.001); B0[2] ~ dnorm(0,0.001)
B2[1] ~ dnorm(0,1000)  # Plausibly exogeneous (informative prior)
B2[2] ~ dnorm(0,0.001); B3[1] ~ dnorm(0,0.001)
...
...
m4_8[2] ~ dnorm(0,0.001); m4_9[2] ~ dnorm(0,0.001)
m4_10[2] ~ dnorm(0,0.001)
}", OperaModel.jags")

rjags.model.opera <- jags.model(OperaModel.jags", data=list(Nlevel1=Nlevel1,
↪  Nlevel2=Nlevel2, home_id=data$home_id, =data$Z, R=data$R1, sppb2=data$sppb2,
↪  sex=data$sex, place=data$place, size=data$size, hometype1=data$hometype1,
↪  hometype2=data$hometype2, antidep0=data$antidep0, age=data$age, sppb0=data$sppb0,
↪  mmse0=data$mmse0), n.chains=n.chains)
rjags.par.opera <- c("A0","B0[1]","B0[2]","B2[1]","B2[2]", "sigma_a", "sigma_b",
↪  "sigma_sppb2")
rjags.sim.opera <- coda.samples(rjags.model.opera, rjags.par.opera, n.burn=n.burnin,
↪  n.iter=n.iter, thin=n.thin)
summary(rjags.sim.opera)

# 2. Assuming no ER and level-2 variance heterogeneity across trial groups and use of
↪  level-2 Sigma half-Cauchy prior
write("model {for (i in 1:Nlevel1) {

# Level-1 missing covariates model specification in increasing order of missingness
antidep0[i]~  dbern(pi_antidep0[i,C[i]])
pi_antidep0[i,1]    <- ilogit(m1[1] + m1_b1[home_id[i]] + m1_2[1]*Z[i] + m1_3[1]*sex[i]
↪  + m1_4[1]*place[i] + m1_5[1]*size[i] + m1_6[1]*hometype1[i] + m1_7[1]*hometype2[i])
pi_antidep0[i,2]    <- ilogit(m1[2] + m1_b2[home_id[i]] + m1_2[2]*Z[i] + m1_3[2]*sex[i]
↪  + m1_4[2]*place[i] + m1_5[2]*size[i] + m1_6[2]*hometype1[i] + m1_7[2]*hometype2[i])

age[i]   ~  dnorm(mu_age[i,C[i]],tau_age)
mu_age[i,1]<- m2[1] + m2_b1[home_id[i]] + m2_2[1]*Z[i] + m2_3[1]*sex[i] +
↪  m2_4[1]*place[i] + m2_5[1]*size[i] + m2_6[1]*hometype1[i] + m2_7[1]*hometype2[i] +
↪  m2_8[1]*antidep0[i]
mu_age[i,2]<- m2[2] + m2_b2[home_id[i]] + m2_2[2]*Z[i] + m2_3[2]*sex[i] +
↪  m2_4[2]*place[i] + m2_5[2]*size[i] + m2_6[2]*hometype1[i] + m2_7[2]*hometype2[i] +
↪  m2_8[2]*antidep0[i]

sppb0[i] ~  dnorm(mu_sppb0[i,C[i]],tau_sppb0)
mu_sppb0[i,1]       <- m3[1] + m3_b1[home_id[i]] + m3_2[1]*Z[i] + m3_3[1]*sex[i] +
↪  m3_4[1]*place[i] + m3_5[1]*size[i] + m3_6[1]*hometype1[i] + m3_7[1]*hometype2[i] +
↪  m3_8[1]*antidep0[i] + m3_9[1]*age[i]
mu_sppb0[i,2]       <- m3[2] + m3_b2[home_id[i]] + m3_2[2]*Z[i] + m3_3[2]*sex[i] +
↪  m3_4[2]*place[i] + m3_5[2]*size[i] + m3_6[2]*hometype1[i] + m3_7[2]*hometype2[i] +
↪  m3_8[2]*antidep0[i] + m3_9[2]*age[i]

mmse0[i] ~  dnorm(mu_mmse0[i,C[i]],tau_mmse0)
mu_mmse0[i,1]       <- m4[1] + m4_b1[home_id[i]] + m4_2[1]*Z[i] + m4_3[1]*sex[i] +
↪  m4_4[1]*place[i] + m4_5[1]*size[i] + m4_6[1]*hometype1[i] + m4_7[1]*hometype2[i] +
↪  m4_8[1]*antidep0[i] + m4_9[1]*age[i] + m4_10[2]*sppb0[i]
mu_mmse0[i,2]       <- m4[2] + m4_b2[home_id[i]] + m4_2[2]*Z[i] + m4_3[2]*sex[i] +
↪  m4_4[2]*place[i] + m4_5[2]*size[i] + m4_6[2]*hometype1[i] + m4_7[2]*hometype2[i] +
↪  m4_8[2]*antidep0[i] + m4_9[2]*age[i] + m4_10[2]*sppb0[i]

# Level-1 outcome model specification
sppb2[i] ~  dnorm(mu_sppb2[i,C[i]],tau_sppb2)
mu_sppb2[i,1]       <- B0 + b_Z0[home_id[i]]*(1-Z[i]) + b_Z1[home_id[i]]*Z[i] +
↪  B2[1]*Z[i] + B3[1]*sex[i] + B4[1]*place[i] + B5[1]*size[i] + B6[1]*hometype1[i] +
↪  B7[1]*hometype2[i] + B8[1]*antidep0[i] + B9[1]*age[i] + B10[1]*sppb0[i] +
↪  B11[1]*mmse0[i]
```

```
mu_sppb2[i,2]       <- B0 + b_Z0[home_id[i]]*(1-Z[i]) + b_Z1[home_id[i]]*Z[i] +
↪  B2[2]*Z[i] + B3[2]*sex[i] + B4[2]*place[i] + B5[2]*size[i] + B6[2]*hometype1[i] +
↪  B7[2]*hometype2[i] + B8[2]*antidep0[i] + B9[2]*age[i] + B10[2]*sppb0[i] +
↪  B11[2]*mmse0[i]

# Level-1 compliance class model specification
R[i]     ~  dbern(pi[i])
pi[i]    <- ilogit(A[i] + A3*sex[i] + A4*place[i] + A5*size[i] + A6*hometype1[i] +
↪  A7*hometype2[i])
A[i]     <- A0 + a[home_id[i]]
C[i]     ~  dcat(p[i,]); p[i,2] <- pi[i]; p[i,1] <- 1-pi[i]
}

for (j in 1:Nlevel2) {
# Level-2 random-effects distribution
b_Z0[j] ~ dnorm(0,tau_b_Z0); b_Z1[j] ~ dnorm(0,tau_b_Z1)
a[j] ~ dnorm(0,tau_a)
m1_b1[j] ~ dnorm(0,tau_m1_b); m1_b2[j] ~ dnorm(0,tau_m1_b)
m2_b1[j] ~ dnorm(0,tau_m2_b); m2_b2[j] ~ dnorm(0,tau_m2_b)
m3_b1[j] ~ dnorm(0,tau_m3_b); m3_b2[j] ~ dnorm(0,tau_m3_b)
m4_b1[j] ~ dnorm(0,tau_m4_b); m4_b2[j] ~ dnorm(0,tau_m4_b)
}

# Priors specification for random effects
tau_b_Z0 <- 1/pow(sigma_b_Z0,2)
sigma_b_Z0 ~ dt(0,pow(100,-2),1)T(0,)
tau_b_Z1 <- 1/pow(sigma_b_Z1,2)
sigma_b_Z1  ~  dt(0,pow(100,-2),1)T(0,)
tau_a    <- 1/pow(sigma_a,2)
sigma_a  ~  dt(0,pow(100,-2),1)T(0,)
tau_m1_b <- 1/pow(sigma_m1_b,2)
sigma_m1_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m2_b <- 1/pow(sigma_m2_b,2)
sigma_m2_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m3_b <- 1/pow(sigma_m3_b,2)
sigma_m3_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m4_b <- 1/pow(sigma_m4_b,2)
sigma_m4_b  ~  dt(0,pow(100,-2),1)T(0,)

# Priors specification for outcome distribution
sigma_sppb2 <- 1/sqrt(tau_sppb2)
tau_sppb2   ~  dgamma(0.001,0.001)
sigma_age<- 1/sqrt(tau_age)
tau_age  ~  dgamma(0.001,0.001)
sigma_sppb0         <- 1/sqrt(tau_sppb0)
tau_sppb0~  dgamma(0.001,0.001)
sigma_mmse0         <- 1/sqrt(tau_mmse0)
tau_mmse0~  dgamma(0.001,0.001)

# Priors specification for outcome model
A0 ~ dnorm(0,0.001); A3 ~ dnorm(0,0.001); A4 ~ dnorm(0,0.001)
A5 ~ dnorm(0,0.001); A6 ~ dnorm(0,0.001); A7 ~ dnorm(0,0.001)
B0 ~ dnorm(0,0.001)
B2[1] ~ dnorm(0,1000)  # Plausibly exogeneous (informative prior)
B2[2] ~ dnorm(0,0.001); B3[1] ~ dnorm(0,0.001)
...
...
m4_8[2] ~ dnorm(0,0.001); m4_9[2] ~ dnorm(0,0.001)
m4_10[2] ~ dnorm(0,0.001)
}", "OperaModel.jags")

rjags.model.opera <- jags.model(OperaModel.jags", data=list(Nlevel1=Nlevel1,
↪  Nlevel2=Nlevel2, home_id=data$home_id, Z=data$Z, R=data$R1, sppb2=data$sppb2,
↪  sex=data$sex, place=data$place, size=data$size, hometype1=data$hometype1,
↪  hometype2=data$hometype2, antidep0=data$antidep0, age=data$age, sppb0=data$sppb0,
↪  mmse0=data$mmse0), n.chains=n.chains)
rjags.par.opera <- c("A0","B0","B2[1]","B2[2]","sigma_a","sigma_b_Z0", "sigma_b_Z1",
↪  "sigma_sppb2")
rjags.sim.opera <- coda.samples(rjags.model.opera, rjags.par.opera, n.burn=n.burnin,
↪  n.iter=n.iter, thin=n.thin)
summary(rjags.sim.opera)


### 3. Assuming no ER and level-2 variance heterogeneity across adherence classes
# 3.1. Adjustment for baseline covariates with missing values and use of level-2 Sigma
↪  half-Cauchy prior
```

```
write("model {for (i in 1:Nlevel1) {

# Level-1 missing covariates model specification in increasing order of missingness
antidep0[i]          ~  dbern(pi_antidep0[i,C[i]])
pi_antidep0[i,1]    <- ilogit(m1[1] + m1_b1[home_id[i]] + m1_2[1]*Z[i] + m1_3[1]*sex[i]
↪  + m1_4[1]*place[i] + m1_5[1]*size[i] + m1_6[1]*hometype1[i] + m1_7[1]*hometype2[i])
pi_antidep0[i,2]    <- ilogit(m1[2] + m1_b2[home_id[i]] + m1_2[2]*Z[i] + m1_3[2]*sex[i]
↪  + m1_4[2]*place[i] + m1_5[2]*size[i] + m1_6[2]*hometype1[i] + m1_7[2]*hometype2[i])

age[i]               ~  dnorm(mu_age[i,C[i]],tau_age)
mu_age[i,1]         <- m2[1] + m2_b1[home_id[i]] + m2_2[1]*Z[i] + m2_3[1]*sex[i] +
↪  m2_4[1]*place[i] + m2_5[1]*size[i] + m2_6[1]*hometype1[i] + m2_7[1]*hometype2[i] +
↪  m2_8[1]*antidep0[i]
mu_age[i,2]         <- m2[2] + m2_b2[home_id[i]] + m2_2[2]*Z[i] + m2_3[2]*sex[i] +
↪  m2_4[2]*place[i] + m2_5[2]*size[i] + m2_6[2]*hometype1[i] + m2_7[2]*hometype2[i] +
↪  m2_8[2]*antidep0[i]

sppb0[i]             ~  dnorm(mu_sppb0[i,C[i]],tau_sppb0)
mu_sppb0[i,1]       <- m3[1] + m3_b1[home_id[i]] + m3_2[1]*Z[i] + m3_3[1]*sex[i] +
↪  m3_4[1]*place[i] + m3_5[1]*size[i] + m3_6[1]*hometype1[i] + m3_7[1]*hometype2[i] +
↪  m3_8[1]*antidep0[i] + m3_9[1]*age[i]
mu_sppb0[i,2]       <- m3[2] + m3_b2[home_id[i]] + m3_2[2]*Z[i] + m3_3[2]*sex[i] +
↪  m3_4[2]*place[i] + m3_5[2]*size[i] + m3_6[2]*hometype1[i] + m3_7[2]*hometype2[i] +
↪  m3_8[2]*antidep0[i] + m3_9[2]*age[i]

mmse0[i]             ~  dnorm(mu_mmse0[i,C[i]],tau_mmse0)
mu_mmse0[i,1]       <- m4[1] + m4_b1[home_id[i]] + m4_2[1]*Z[i] + m4_3[1]*sex[i] +
↪  m4_4[1]*place[i] + m4_5[1]*size[i] + m4_6[1]*hometype1[i] + m4_7[1]*hometype2[i] +
↪  m4_8[1]*antidep0[i] + m4_9[1]*age[i] + m4_10[2]*sppb0[i]
mu_mmse0[i,2]       <- m4[2] + m4_b2[home_id[i]] + m4_2[2]*Z[i] + m4_3[2]*sex[i] +
↪  m4_4[2]*place[i] + m4_5[2]*size[i] + m4_6[2]*hometype1[i] + m4_7[2]*hometype2[i] +
↪  m4_8[2]*antidep0[i] + m4_9[2]*age[i] + m4_10[2]*sppb0[i]

# Level-1 outcome model specification
sppb2[i]             ~  dnorm(mu_sppb2[i,C[i]],tau_sppb2)
mu_sppb2[i,1]       <- B0[1] + b1_C[home_id[i]] + B2[1]*Z[i] + B3[1]*sex[i] +
↪  B4[1]*place[i] + B5[1]*size[i] + B6[1]*hometype1[i] + B7[1]*hometype2[i] +
↪  B8[1]*antidep0[i] + B9[1]*age[i] + B10[1]*sppb0[i] + B11[1]*mmse0[i]
mu_sppb2[i,2]       <- B0[2] + b2_C[home_id[i]] + B2[2]*Z[i] + B3[2]*sex[i] +
↪  B4[2]*place[i] + B5[2]*size[i] + B6[2]*hometype1[i] + B7[2]*hometype2[i] +
↪  B8[2]*antidep0[i] + B9[2]*age[i] + B10[2]*sppb0[i] + B11[2]*mmse0[i]

# Level-1 compliance class model specification
R[i]  ~  dbern(pi[i])
pi[i] <- ilogit(A[i] + A3*sex[i] + A4*place[i] + A5*size[i] + A6*hometype1[i] +
↪  A7*hometype2[i])
A[i]  <- A0 + a[home_id[i]]
C[i] ~ dcat(p[i,]); p[i,2] <- pi[i]; p[i,1] <- 1-pi[i]
}

for (j in 1:Nlevel2) {
# Level-2 random-effects distribution
b1_C[j] ~ dnorm(0,tau_b1_C); b2_C[j] ~ dnorm(0,tau_b2_C)
a[j] ~ dnorm(0,tau_a)
m1_b1[j] ~ dnorm(0,tau_m1_b); m1_b2[j] ~ dnorm(0,tau_m1_b)
m2_b1[j] ~ dnorm(0,tau_m2_b); m2_b2[j] ~ dnorm(0,tau_m2_b)
m3_b1[j] ~ dnorm(0,tau_m3_b); m3_b2[j] ~ dnorm(0,tau_m3_b)
m4_b1[j] ~ dnorm(0,tau_m4_b); m4_b2[j] ~ dnorm(0,tau_m4_b)
}

# Priors specification for random effects
tau_b1_C <- 1/pow(sigma_b1_C,2)
sigma_b1_C ~  dt(0,pow(100,-2),1)T(0,)
tau_b2_C <- 1/pow(sigma_b2_C,2)
sigma_b2_C ~  dt(0,pow(100,-2),1)T(0,)
tau_a <- 1/pow(sigma_a,2)
sigma_a  ~  dt(0,pow(100,-2),1)T(0,)
tau_m1_b <- 1/pow(sigma_m1_b,2)
sigma_m1_b ~  dt(0,pow(100,-2),1)T(0,)
tau_m2_b <- 1/pow(sigma_m2_b,2)
sigma_m2_b ~  dt(0,pow(100,-2),1)T(0,)
tau_m3_b <- 1/pow(sigma_m3_b,2)
sigma_m3_b ~  dt(0,pow(100,-2),1)T(0,)
tau_m4_b <- 1/pow(sigma_m4_b,2)
sigma_m4_b ~  dt(0,pow(100,-2),1)T(0,)
```

```
# Priors specification for outcome distribution
sigma_sppb2 <- 1/sqrt(tau_sppb2); tau_sppb 2~ dgamma(0.001,0.001)
sigma_age<- 1/sqrt(tau_age); tau_age ~ dgamma(0.001,0.001)
sigma_sppb0<- 1/sqrt(tau_sppb0); tau_sppb0~ dgamma(0.001,0.001)
sigma_mmse0<- 1/sqrt(tau_mmse0); tau_mmse0~ dgamma(0.001,0.001)

# Priors specification for outcome model
A0 ~ dnorm(0,0.001); A3 ~ dnorm(0,0.001)
A4 ~ dnorm(0,0.001); A5 ~ dnorm(0,0.001)
A6 ~ dnorm(0,0.001); A7 ~ dnorm(0,0.001)
B0[1] ~ dnorm(0,0.001); B0[2] ~  dnorm(0,0.001)
B2[1] ~ dnorm(0,1000) # Plausibly exogeneous
B2[2] ~ dnorm(0,0.001); B3[1] ~ dnorm(0,0.001)
...
...
m4_8[2] ~ dnorm(0,0.001); m4_9[2] ~ dnorm(0,0.001)
m4_10[2] ~ dnorm(0,0.001)
}", "OperaModel.jags")

rjags.model.opera <- jags.model("OperaModel.jags", data=list(Nlevel1=Nlevel1,
↪  Nlevel2=Nlevel2, home_id=data$home_id, Z=data$Z, R=data$R1, sppb2=data$sppb2,
↪  sex=data$sex, place=data$place, size=data$size, hometype1=data$hometype1,
↪  hometype2=data$hometype2, antidep0=data$antidep0, age=data$age, sppb0=data$sppb0,
↪  mmse0=data$mmse0), n.chains=n.chains)
rjags.par.opera <- c("A0","B0[1]","B0[2]","B2[1]", "B2[2]", "sigma_a", "sigma_b1_C",
↪  "sigma_b2_C", "sigma_sppb2")
rjags.sim.opera <- coda.samples(rjags.model.opera, rjags.par.opera, n.burn=n.burnin,
↪  n.iter=n.iter, thin=n.thin, na.rm=FALSE)
summary(rjags.sim.opera)

### 4. Assuming no ER and level-1 variance heterogeneity across adherence classes
# 4.1. Adjustment for baseline covariates with missing values and use of level-2 Sigma
↪  half-Cauchy prior
write("model {for (i in 1:Nlevel1) {

# Level-1 missing covariates model specification in increasing order of missingness
antidep0[i]          ~  dbern(pi_antidep0[i,C[i]])
pi_antidep0[i,1]     <- ilogit(m1[1] + m1_b1[home_id[i]] + m1_2[1]*Z[i] + m1_3[1]*sex[i]
↪  + m1_4[1]*place[i] + m1_5[1]*size[i] + m1_6[1]*hometype1[i] + m1_7[1]*hometype2[i])
pi_antidep0[i,2]     <- ilogit(m1[2] + m1_b2[home_id[i]] + m1_2[2]*Z[i] + m1_3[2]*sex[i]
↪  + m1_4[2]*place[i] + m1_5[2]*size[i] + m1_6[2]*hometype1[i] + m1_7[2]*hometype2[i])

age[i]               ~  dnorm(mu_age[i,C[i]],tau_age)
mu_age[i,1]          <- m2[1] + m2_b1[home_id[i]] + m2_2[1]*Z[i] + m2_3[1]*sex[i] +
↪  m2_4[1]*place[i] + m2_5[1]*size[i] + m2_6[1]*hometype1[i] + m2_7[1]*hometype2[i] +
↪  m2_8[1]*antidep0[i]
mu_age[i,2]          <- m2[2] + m2_b2[home_id[i]] + m2_2[2]*Z[i] + m2_3[2]*sex[i] +
↪  m2_4[2]*place[i] + m2_5[2]*size[i] + m2_6[2]*hometype1[i] + m2_7[2]*hometype2[i] +
↪  m2_8[2]*antidep0[i]

sppb0[i]             ~  dnorm(mu_sppb0[i,C[i]],tau_sppb0)
mu_sppb0[i,1]        <- m3[1] + m3_b1[home_id[i]] + m3_2[1]*Z[i] + m3_3[1]*sex[i] +
↪  m3_4[1]*place[i] + m3_5[1]*size[i] + m3_6[1]*hometype1[i] + m3_7[1]*hometype2[i] +
↪  m3_8[1]*antidep0[i] + m3_9[1]*age[i]
mu_sppb0[i,2]        <- m3[2] + m3_b2[home_id[i]] + m3_2[2]*Z[i] + m3_3[2]*sex[i] +
↪  m3_4[2]*place[i] + m3_5[2]*size[i] + m3_6[2]*hometype1[i] + m3_7[2]*hometype2[i] +
↪  m3_8[2]*antidep0[i] + m3_9[2]*age[i]

mmse0[i]             ~  dnorm(mu_mmse0[i,C[i]],tau_mmse0)
mu_mmse0[i,1]        <- m4[1] + m4_b1[home_id[i]] + m4_2[1]*Z[i] + m4_3[1]*sex[i] +
↪  m4_4[1]*place[i] + m4_5[1]*size[i] + m4_6[1]*hometype1[i] + m4_7[1]*hometype2[i] +
↪  m4_8[1]*antidep0[i] + m4_9[1]*age[i] + m4_10[2]*sppb0[i]
mu_mmse0[i,2]        <- m4[2] + m4_b2[home_id[i]] + m4_2[2]*Z[i] + m4_3[2]*sex[i] +
↪  m4_4[2]*place[i] + m4_5[2]*size[i] + m4_6[2]*hometype1[i] + m4_7[2]*hometype2[i] +
↪  m4_8[2]*antidep0[i] + m4_9[2]*age[i] + m4_10[2]*sppb0[i]

# Level-1 outcome model specification
sppb2[i] ~ dnorm(mu_sppb2[i,C[i]],tau_sppb2[C[i]])
mu_sppb2[i,1] <- B0[1] + b1_C[home_id[i]] + B2[1]*Z[i] + B3[1]*sex[i] + B4[1]*place[i] +
↪  B5[1]*size[i] + B6[1]*hometype1[i] + B7[1]*hometype2[i] + B8[1]*antidep0[i] +
↪  B9[1]*age[i] + B10[1]*sppb0[i] + B11[1]*mmse0[i]
mu_sppb2[i,2] <- B0[2] + b2_C[home_id[i]] + B2[2]*Z[i] + B3[2]*sex[i] + B4[2]*place[i] +
↪  B5[2]*size[i] + B6[2]*hometype1[i] + B7[2]*hometype2[i] + B8[2]*antidep0[i] +
↪  B9[2]*age[i] + B10[2]*sppb0[i] + B11[2]*mmse0[i]

# Level-1 compliance class model specification
```

```
R[i]  ~  dbern(pi[i])
pi[i] <- ilogit(A[i] + A3*sex[i] + A4*place[i] + A5*size[i] + A6*hometype1[i] +
↪  A7*hometype2[i])
A[i]  <- A0 + a[home_id[i]]
C[i]  ~  dcat(p[i,]); p[i,2]<- pi[i]; p[i,1]<- 1-pi[i]
}

for (j in 1:Nlevel2) {
# Level-2 random-effects distribution
b1_C[j] ~ dnorm(0,tau_b_C); b2_C[j] ~ dnorm(0,tau_b_C)
a[j] ~ dnorm(0,tau_a)
m1_b1[j] ~ dnorm(0,tau_m1_b); m1_b2[j] ~ dnorm(0,tau_m1_b)
m2_b1[j] ~ dnorm(0,tau_m2_b); m2_b2[j] ~ dnorm(0,tau_m2_b)
m3_b1[j] ~ dnorm(0,tau_m3_b); m3_b2[j] ~ dnorm(0,tau_m3_b)
m4_b1[j] ~ dnorm(0,tau_m4_b); m4_b2[j] ~ dnorm(0,tau_m4_b)
}

# Priors specification for random effects
tau_b_C     <- 1/pow(sigma_b_C,2)
sigma_b_C   ~  dt(0,pow(100,-2),1)T(0,)
tau_a       <- 1/pow(sigma_a,2)
sigma_a     ~  dt(0,pow(100,-2),1)T(0,)
tau_m1_b    <- 1/pow(sigma_m1_b,2)
sigma_m1_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m2_b    <- 1/pow(sigma_m2_b,2)
sigma_m2_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m3_b    <- 1/pow(sigma_m3_b,2)
sigma_m3_b  ~  dt(0,pow(100,-2),1)T(0,)
tau_m4_b    <- 1/pow(sigma_m4_b,2)
sigma_m4_b  ~  dt(0,pow(100,-2),1)T(0,)

# Priors specification for outcome distribution
sigma_sppb2[1] <- 1/sqrt(tau_sppb2[1])
tau_sppb2[1]   ~  dgamma(0.001,0.001)
sigma_sppb2[2] <- 1/sqrt(tau_sppb2[2])
tau_sppb2[2]   ~  dgamma(0.001,0.001)
sigma_age      <- 1/sqrt(tau_age)
tau_age        ~  dgamma(0.001,0.001)
sigma_sppb0    <- 1/sqrt(tau_sppb0)
tau_sppb0      ~  dgamma(0.001,0.001)
sigma_mmse0    <- 1/sqrt(tau_mmse0)
tau_mmse0      ~  dgamma(0.001,0.001)

# Priors specification for outcome model
A0 ~ dnorm(0,0.001); A3 ~ dnorm(0,0.001)
A4 ~ dnorm(0,0.001); A5 ~ dnorm(0,0.001)
A6 ~ dnorm(0,0.001); A7 ~ dnorm(0,0.001)
B0[1] ~ dnorm(0,0.001); B0[2] ~  dnorm(0,0.001)
B2[1] ~ dnorm(0,1000)  # Plausibly exogeneous
B2[2] ~ dnorm(0,0.001); B3[1] ~ dnorm(0,0.001)
...
...
m4_8[2] ~ dnorm(0,0.001); m4_9[2] ~ dnorm(0,0.001)
m4_10[2] ~ dnorm(0,0.001)
}", "OperaModel.jags")

rjags.model.opera <- jags.model(OperaModel.jags", data=list(Nlevel1=Nlevel1,
↪  Nlevel2=Nlevel2, home_id=data$home_id, Z=data$Z, R=data$R1, sppb2=data$sppb2,
↪  sex=data$sex, place=data$place, size=data$size, hometype1=data$hometype1,
↪  hometype2=data$hometype2, antidep0=data$antidep0, age=data$age, sppb0=data$sppb0,
↪  mmse0=data$mmse0), n.chains=n.chains)
rjags.par.opera <- c("A0","B0[1]","B0[2]","B2[1]", "B2[2]", "sigma_a", "sigma_b_C",
↪  "sigma_sppb2[1]", "sigma_sppb2[2]")
rjags.sim.opera <- coda.samples(rjags.model.opera, rjags.par.opera, n.burn=n.burnin,
↪  n.iter=n.iter, thin=n.thin, na.rm=FALSE)
summary(rjags.sim.opera)
```